Journal of Korea Multimedia Society Vol. 22, No. 11, November 2019(pp. 1223-1232)
https://doi.org/10.9717/kmms.2019.22.11.1223

Bi-GRU 이미지 캡션의 서술 성능 향상을 위한 Parallel Injection 기법 연구

이준희[†], 이수환[††], 태수호[†††], 서동환[††††]

Parallel Injection Method for Improving Descriptive Performance of Bi-GRU Image Captions

Jun Hee Lee[†], Soo Hwan Lee[††], Soo Ho Tae[†††], Dong Hoan Seo[††††]

ABSTRACT

The injection is the input method of the image feature vector from the encoder to the decoder. Since the image feature vector contains object details such as color and texture, it is essential to generate image captions. However, the bidirectional decoder model using the existing injection method only inputs the image feature vector in the first step, so image feature vectors of the backward sequence are vanishing. This problem makes it difficult to describe the context in detail. Therefore, in this paper, we propose the parallel injection method to improve the description performance of image captions. The proposed Injection method fuses all embeddings and image vectors to preserve the context. Also, We optimize our image caption model with Bidirectional Gated Recurrent Unit (Bi-GRU) to reduce the amount of computation of the decoder. To validate the proposed model, experiments were conducted with a certified image caption dataset, demonstrating excellence in comparison with the latest models using BLEU and METEOR scores. The proposed model improved the BLEU score up to 20.2 points and the METEOR score up to 3.65 points compared to the existing caption model.

Key words: Image Caption, Parallel Injection, Bi-GRU, Injection Method

1. 서 론

이미지의 내용을 이해하고 서술하는 작업은 영상의학, 법영상, 무인감시 등의 다양한 분야에서 필수적이다. 이 분야들은 하드웨어 및 소프트웨어의 발달로 인해 부분적인 자동화[1-4]가 진행되고 있지만 정보 간 관계를 이해하는 분석 영역에서는 수작업에 의존한다. 이 문제의 완전한 자동화를 위해서 기계가 능동적으로 이미지의 상황을 이해하고 문장으로 표현하는 이미지 캡션 기술이 활발하게 연구되고 있다. 그러나 이 기법은 객체들의 겹침, 각도, 조명의 변화 등 다양한 환경적인 영향에 의해 오차가 발생하기

※ Corresponding Author : Dong Hoan Seo, Address: (49112) Engineering Building1 543 Taejong-ro 727, Youngdo-gu, Busan, Korea, TEL : +82-51-410-4412, FAX : +82-51-410-4412, E-mail : dhseo@kmou.ac.kr
Receipt date : May 31, 2019, Revision date : Sep. 27, 2019
Approval date : Oct. 22, 2019
[†] Dept. of Electrical and Electronics Engineering, Korea Maritime and Ocean University
(E-mail : ljh9961@naver.com)
[††] Dept. of Electrical and Electronics Engineering, Korea Maritime and Ocean University
(E-mail : config5246@naver.com)
[†††] Dept. of Electrical and Electronics Engineering, Korea Maritime and Ocean University
(E-mail : heavenroute@naver.com)
[††††] Div. of Electronics and Electrical Information Engineering, Korea Maritime and Ocean University
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No.2016R1D1A1B03934812).

때문에 사람과 대등한 수준의 결과를 만들기 위해선 다양한 과제들이 존재한다[5-10].

최근의 이미지 캡션 기술은 주어진 이미지를 문장으로 변형하기 위해 이미지에서 특징을 획득하는 Convolutional Neural Network(CNN)로 구성된 인코더와 단어를 문장으로 재배치하는 Recurrent Neural Network(RNN)로 구성된 디코더를 가지는 인코더-디코더 모델을 적용한다. 이러한 접근법은 이미지의 객체를 바탕으로 문장을 생성 가능하지만 RNN 기반 디코더의 한계인 경사 소실 문제와 이미지 특징 벡터 소실 문제로 인해 문장의 길이가 길어질수록 문장 구조가 파괴되기 쉽다. 문장 구조의 파괴는 단어나 문장 배치의 문제인 구문적 오류는 없으나 원래 의미와 다른 결과가 나오는 문맥적 오류를 발생시킨다. 이러한 오류는 이미지의 상황에 정확한 전달을 어렵게 만들기 때문에 기계의 능동적 상황 이해와는 차이가 있다.

따라서 본 논문에서는 이미지 특징 벡터의 소멸로 인한 문장 구조의 파괴와 이미지의 내용과는 전혀 다른 문장이 생성되는 문맥적인 오류를 최소화하기 위한 캡션 모델을 제안한다. 제안하는 이미지 캡션 모델은 기존의 연구들과 달리 임베딩 레이어에 계속 이미지 특징 벡터를 입력하는 Parallel Injection 방식을 사용하며, 양방향 구조를 통해 이미지 특징 벡터 소실 문제를 최소화한다. 또한 경사 소실 문제에 강건하고 적은 연산을 필요로 하는 GRU를 적용하여 모든 RNN 노드에 이미지 특징 벡터가 입력되어 경사 소실로 인한 이미지 특징 벡터 소실을 막을 수 있다. 본 모델은 이를 통해 이미지 캡션 모델에서 발생하는 구문적 오류를 줄여 정확도를 높이며, 이미지에 대한 상위 수준의 의미전달이 가능한 상세한 캡션이 가능하다. 또한 이전 단어뿐만 아니라 이후에 나오는 단어도 학습하여 배치를 고려하므로 문맥적 오류가 적은 캡션 표현이 가능하다.

2. 관련 연구

최근의 높은 성과를 내는 딥러닝 기반 이미지 캡션 모델은 인코더-디코더 구조로 설계된다. 인코더-디코더 모델은 이미지 내의 객체를 인식하는 객체 인식 기법과 단어를 통해 문장으로 재배치하는 자연어 처리 기법이 결합된 기술이다.

2.1 이미지 객체 인식 기법

일반적인 CNN은 이미지를 입력받는 네트워크로써 객체 분류 및 검출의 분야에서 우수한 성능을 보이고 있다. ImageNet에서 우수한 분류 성능을 보인 VGGNet[11], Inception V3[12], ResNet[13]과 같은 모델뿐만 아니라 최근에는 실시간 객체 검출이 가능한 YOLO[14]와 정밀한 객체 검출이 가능한 Faster R-CNN[15]과 같은 다양한 모델이 존재한다. 이미지 캡션 모델에서 인코더로 널리 사용하는 Inception V3는 2014년 ImageNet Large Scale Visual Recognition Competition(IRSVRC)에서 좋은 성능을 보인 모델이다. 딥러닝은 레이어가 깊고 넓을수록 학습 성능이 좋아지지만 기존의 다양한 CNN모델들은 과적합, 경사 소실 문제와 같은 현상으로 인해 레이어를 적층만 해서는 학습 성능이 오히려 저하된다. Inception V3는 레이어의 적층으로 발생하는 다양한 문제를 해결하기 위해 레이어에 Inception 모듈이라는 새로운 개념의 레이어를 적용함으로써 레이어 적층으로 인해 발생하는 과적합을 방지하고 역전파 과정에서 가중치 갱신이 잘 되도록 함으로써 경사 소실 문제를 해결하여 최고의 성능을 보였다.

2.2 이미지 캡션 생성 기법

추출된 객체를 문장으로 만드는 디코더는 이미지 캡션 모델의 핵심으로써 데이터 세트로부터 획득한 단어 벡터를 통해 인코더에서 획득된 이미지 캡션 순서를 재배치한다. 하지만 RNN의 동작 과정에서 발생하는 이미지 특징 벡터 소멸 문제로 인해 객체가 캡션에 반영되지 않는 것을 해결하기 위해 다양한 연구가 진행되고 있다.

O. Vinyals et al. [16]은 인코더-디코더 구조로, CNN의 이미지 특징을 사용하여 객체를 검출하고, Long Short-Term Memory(LSTM)을 사용하여 캡션을 생성함으로써 경사 소실 문제(Vanishing Gradient Problem)를 해결하였다. 하지만 생성되는 캡션에는 객체들의 속성이나 관계와 같은 상위 수준의 의미 정보가 추출되지 않는 단점이 있다. K. Xu et al. [17]는 앞선 인코더-디코더 모델에서 특정 부분에 관심을 집중하는지를 선택하는 어텐션 기법을 사용하여 이미지 캡션을 제안하였다. 이를 통해 캡션과 이미지의 연관성은 향상되었으나 생성되는 캡션이 단

순하여 세밀한 정보를 전달하지 못하는 한계가 있다.

최근 J. Mao et al. [18]은 이미지와 단어 벡터를 결합하는 멀티모달 레이어를 적용한 이미지 캡션 모델을 제안하였다. 이 모델은 인코더의 이미지 특징을 멀티모달 레이어로 피드백하여 생성되는 캡션과 이미지의 연관성이 향상되었으나 문장을 생성하는 RNN 과정의 뒤에 삽입되어 효과가 제한적이다. 또한 기본적인 RNN을 사용하기 때문에 경사 소실 문제에 취약하며, 단일 방향으로 동작하는 RNN을 사용하여 최초의 단계에서만 이미지 특징을 입력하기 때문에 문장의 길이에 따라 이미지 특징 벡터 소멸 문제를 가지고 있다. 이를 개선하여 J. Guan et al. [19]은 멀티모달뿐만 아니라 임베딩 레이어도 이미지 특징을 피드백하여 문장의 생성 과정에서 이미지 특징 벡터 소멸을 개선하였다. 또한 Bidirectional Recurrent Neural Network(Bi-RNN) 구조[20]를 바탕의 LSTM 레이어를 통해 양방향의 문장 특성을 고려하였다. 하지만 이 모델은 RNN의 첫 순간에만 이미지 특징 벡터가 입력되는 Prefix Injection 방식을 적용했기 때문에 역방향 RNN에는 이미지 특징 벡터가 전달되지 않아 Bi-RNN 구조에 최적화 되지 않았다.

3. 제안한 방법

본 연구는 더블 임베딩 및 멀티모달을 적용한 m-RNN[18]과 Bi-RNN 구조를 사용한 Repeated review[19]의 문제점을 개선한다. 제안한 모델은 LSTM을 간소화한 Gated Recurrent Unit(GRU)[21]를 적용하고 양방향 RNN 구조에 최적화된 임베딩 및 모달 레이어에 이미지 특징 벡터를 입력해주는 Parallel Injection 방식을 통해 구문적 오류뿐만 아니라 문맥적 오류를 최소화하는 이미지 캡션 모델을 설계하였다.

3.1 전체 구성

제안하는 모델의 구성은 인코더, 임베딩, Bi-RNN과 멀티모달 레이어로 구성된다. Fig. 1은 제안된 모델의 전체 구성을 나타낸다. Fig. 1 좌측 하단의 인코더는 이미지를 특징 벡터로 만들어 이중 임베딩 레이어의 정보와 결합한다. Embedding Ⅰ은 데이터 세트에 포함된 단어를 원-핫 인코딩을 통해 벡터화하고 워드 임베딩 과정을 통해 단어 벡터를 생성한다. 임베딩 레이어를 Embedding Ⅱ까지 확장하여 단일 레이어에 비해 문장 표현력이 향상된다. Fig. 1의 Parallel Injection 부분에서 기존의 다른 연구와 달리 임베딩 레이어는 인코더의 결과를 첫 단계의 임베딩에 입력해주는 과정을 확장하여 모든 임베딩 순간에 삽입함으로써 디코더에서 발생하는 이미지 특징의 감소를 방지한다. Embedding Ⅱ의 단어와 이미지 특징이 융합된 벡터는 Bi-RNN으로 구성된 디코더의 입력으로 사용된다. 디코더는 양방향 특징을 고려한 학습이 가능한 Bi-RNN 구조를 적용하여 문장의 순서를 출력한다. 또한 디코더는 최적화를 위해 Bi-RNN의 연산량을 최대한 줄이고 경사 소실 문제에 강인한 GRU를 적용한다. 멀티모달 레이어는 인코더에서 추출되는 이미지 특징 벡터, Embedding Ⅱ의 단어 벡터와 Bi-RNN에서 획득하는 문장 벡터를 단일 특징 공간에 표현하고 캡션을 생성한다.

3.2 인코더

이미지 캡션은 이미지에 대한 설명을 생성하기 때문에 인코더에서 이미지의 특징 추출이 요구된다. Fig. 1의 좌측 인코더가 이 부분을 나타낸다. 제안하는 모델의 인코더는 ImageNet에서 우수한 성능을 보인 Inception V3[12]를 사용한다. 인코더는 기존의 ImageNet 데이터세트를 이용하여 선학습 되어있는 가중치지도를 이용한다. 다층 CNN의 동작 특성은 상위 레이어가 이미지의 전체적인 정보를 가지고 있으며, 이미지의 부분적인 공간 영역에 분포하는 객체 정보는 하위 레이어에 존재한다. 따라서 인코더는 하위 레이어의 특징을 추출할 필요가 있다. 제안하는 이미지 캡션 모델은 부분적인 공간 영역에 대한 객체 및 이미지의 정보가 포함되어 있는 완전연결 레이어에서 이미지 특징을 추출하여 이미지 캡션에 사용한다. 완전연결 레이어에서 캡션에 사용되는 이미지 특징을 추출하기 위해서는 이미지에 주어진 캡션과 어휘의 크기를 사용하여 이미지와 대응하는 차원으로 캡션을 정리하는 과정이 필요하다. 캡션을 정리하는 과정은 다음 식과 같다.

$$y = y_1, \cdots, y_c \in R^K \tag{1}$$

여기서 K는 단어의 수이고 c는 캡션의 길이를 의미한다. 정리된 캡션을 이용해 이미지의 특징 벡터를 추출하는 과정에서는 CNN이 사용되며 추출된 각 벡터는 이미지의 일부분에 대응하는 차원의 표현으로

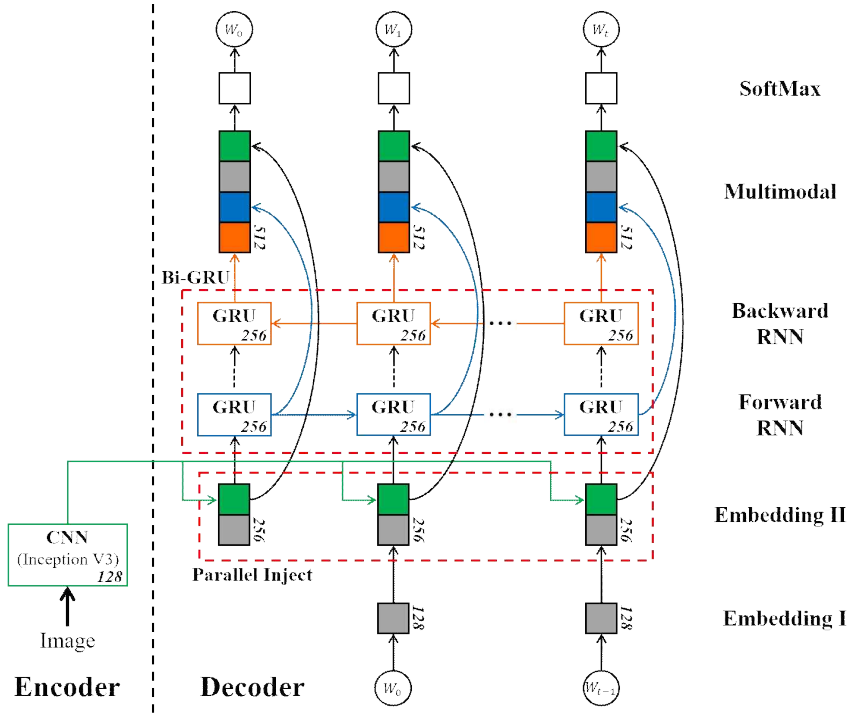

Fig. 1. Proposed Image caption model.

나타난다. 다음 식은 이미지 특징을 추출하는 방식을 나타낸다.

$$a = a_1, \cdots, a_L \in R^D \tag{2}$$

여기서 L은 추출된 일련의 특징 벡터의 개수, D는 차원을 의미하며 이렇게 획득된 이미지 특징 벡터는 Embedding II와 멀티모달 레이어에 입력된다.

3.3 Parallel Injection

Fig. 1 하단의 적색 상자는 본 논문에서 제안하는 Parallel Injection 방식을 표현한 것이다. 이미지 특징 벡터 소멸 문제를 해결하기 위해 기존의 연구들은 임베딩 및 멀티모달 레이어에 이미지 특징 벡터와 단어 특징 벡터를 융합하는 방식을 적용하였다. 삽입하는 순간에 따라 Initial Injection, Multimodal Injection, Prefix Injection 그리고 Parallel Injection 방식이 있다. Initial Injection는 이미지 특징 벡터를 단어 벡터로 변경하여 각 임베딩 레이어에 입력해주는 초기 이미지 캡션 모델에서 사용된 방식으로 y_t가 각 순간에 입력되기 때문에 이전의 단어가 소멸하기 쉽다. Multimodal Injection는 m-RNN에서 제안된 방식으로 모든 멀티모달 레이어에 이미지 특징 벡터 a가 삽입된다. 하지만 문장을 생성하는 RNN 과정 뒤에 삽입되기 때문에 문장 생성에 영향을 줄 수 없다. Prefix Injection 은 repeated review에서 제안된 방식으로 초기 임베딩 레이어에 a를 삽입하여 RNN에 직접 영향을 주는 방식이다. 이 방식을 통해 문장 생성의 정확도는 향상되었으나 역방향 RNN에는 영향을 주지 않기 때문에 캡션 생성 정확도 향상이 제한적이다. 따라서 본 논문에서는 모든 임베딩 레이어에 이미지 특징 벡터 a를 삽입하여 Bi-RNN 구조에 최적화된 Parallel Injection 방식을 사용한다. Parallel Injection 방식은 인코더에서 획득한 2개의 벡터를 융합하여 단어 벡터와 이미지 벡터가 지속적으로 디코더에 주입한다. 이를 통해 임베딩 레이어는 RNN 레이어 동작 과정에서 발생하는 이미지 특징의 소멸을 방지한다. 또한 Multimodal Injection 방식도 동일하게 적용하여 이미지의 세부적인 특징이나 다중 객체가 등장하는 경우, 객체에 대한 설명이 누락되는 경우를 방지한다. 제안하는 모델은 이 두 방식을 통해 이미지에 포함된 작은 객체에 대해서도 문장 표현이 가능하며 전체 이미지의 상황에 대한 고려도

가능하다.

3.4 Bi-GRU

이중 임베딩 기법을 통해 생성되는 이미지와 단어의 융합 벡터는 양방향으로 구성된 GRU의 입력으로 사용된다. 기본적인 RNN이 가지는 경사 소실 문제를 해결하기 위해 GRU는 리셋과 업데이트 게이트로 이루어져 있으며 LSTM을 간소화시켰다. 또한 GRU에 양방향 구조를 적용하여 역방향과 순방향 단어 특징을 참고할 수 있도록 함으로써 현재 시간에서 단어를 생성할 때 전체적인 문맥을 고려한 단어 생성이 가능하다. GRU의 리셋 게이트는 현재 시점의 입력 데이터와 이전 시간의 출력 데이터를 선택한다. 업데이트 게이트는 기억의 정도를 판단한다. 이러한 GRU의 2개의 게이트를 이용하여 현재 셀이 가지는 정보를 도출하며 마지막으로 현재 셀의 정보와 과거 셀의 정보가 합쳐진 데이터가 출력된다. GRU는 다음과 같이 정의된다.

$$r_t = \sigma(W_r \bullet [h_{t-1} \ , \ x_t]) \tag{3}$$

$$z_t = \sigma(W_z \bullet [h_{t-1} \ , \ x_t]) \tag{4}$$

$$\tilde{h}_t = \tanh(W \bullet [r_t \times h_{t-1} \ , \ x_t]) \tag{5}$$

$$h_t = (1-z_z) \times h_{t-1} + z_t \times \tilde{h}_t \tag{6}$$

여기서 W는 각 구간에서 사용되는 가중치값, r_t는 리셋 게이트, z_t는 업데이트 게이트를 의미한다. 또한 $\tilde{h}_t$는 현재 셀이 가지는 정보, h_t는 출력을 의미한다. Bi-GRU의 양방향 동작 특성을 통해 이전 단어뿐만 아니라 전체 문맥에 대한 고려가 가능하여 생성되는 문장이 길어져도 문맥적 오류가 적어진다는 장점을 가진다. 또한 지속적인 GRU 동작 과정에서 발생하는 이미지 특징의 소멸 문제를 Embedding Ⅱ에서 생성되는 융합 벡터를 통해 해결함으로써 기존 이미지 캡션 모델에 비해 월등히 많은 이미지 특징을 유지한다. 이러한 과정을 통해 Bi-GRU에서 생성되는 문장 순서의 특징은 멀티모달 레이어의 최종적인 캡션 문장의 문맥적 의미 향상을 위해 주입된다.

4. 실험 결과 및 고찰

4.1 이미지 캡션을 위한 데이터세트 구성

제안된 이미지 캡션 모델은 공인된 Flickr 8K[22], Flickr 30K[23]와 MS-COCO[24] 데이터세트를 사용하여 학습 및 검증을 진행하였다. 실험은 Flickr에서 추출한 8,000장의 이미지로 구성된 Flickr 8K를 학습 및 검증, 테스트를 진행하기 위해 각각 6,000장, 1,000장, 1,000장을 사용한다. 또한 Flickr 8K의 확장된 데이터세트인 Flickr 30K는 30,000장의 이미지를 제공하며 검증 및 테스트에 사용하기 위해 동일한 비율로 구성한다. 마지막으로 MS-COCO 데이터세트는 82,783개의 교육 이미지와 40,504개의 검증 이미지가 포함되어 있다. 또한 학습을 진행하기 이전에 데이터세트에서 5회 미만으로 등장하는 단어를 포함한 캡션은 학습 및 검증 데이터에서 배제하도록 전처리 과정을 거친다. 이 과정을 통해 빈도가 높은 단어 위주로 캡션을 일반화하여 모델이 정확한 문장을 생성 가능하다.

4.2 Injection 방식별 비교 및 분석

데이터가 재귀 입력되는 RNN의 구조적 특징으로 인해 순차가 늘어남에 따라 경사 소실 이 커지기 때문에 초기 입력된 데이터는 이후의 시간에 사라질 가능성이 높다. 따라서 이미지의 캡션이 증가할수록 경사 소실 문제에 취약해져 문장의 구조 및 캡션이 파괴될 가능성이 높다. 그렇기 때문에 마지막에 생존한 이미지 특징 벡터의 크기가 크면 경사 소실 문제에 강건하다. Fig. 2는 각 방식에 따른 이미지 특징 벡터의 크기를 보여준다.

Initial Injection 방식은 이미지 특징 벡터를 RNN의 초기 상태 벡터로 사용되어 RNN을 초기화 한 후 단어 벡터를 통해 캡션을 생성한다. 따라서 초기 상태 벡터의 크기에 비례하여 이미지 벡터의 크기가 결정되기 때문에 제일 적은 이미지 벡터 크기를 가지는 것을 보여준다.

Prefix Injection 방식은 Repeated review에서 적용된 방식으로 Initial Injection 방식과 달리 단어 벡터와 동일한 크기의 이미지 벡터를 사용하기 때문에 이미지 벡터의 크기가 증가하는 것을 볼 수 있다. 하지만 이 방식의 경우 캡션 생성이 진행될수록 이미지 벡터의 크기가 계속해서 감소되는 것을 확인할 수 있다.

본 모델에서 사용하는 Parallel Injection 방식은 단어 벡터가 입력으로 사용될 때 이미지 벡터도 함께 입력되기 때문에 캡션 생성 과정에서 이미지 벡터가 계속 공급되어 벡터의 크기가 증감을 반복하는 모습

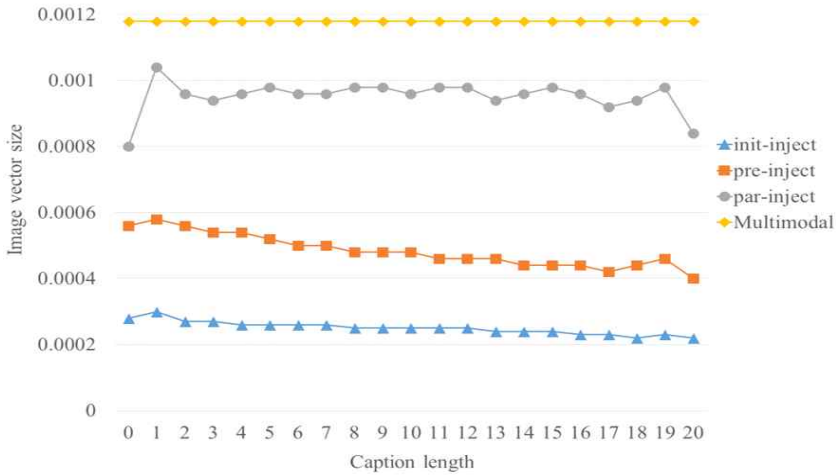

Fig. 2. Decrease of image vector size due to caption generation.

을 보이며 캡션의 길이가 늘어나도 전체적인 벡터의 크기 감소가 다른 방식들에 비해 일정한 것을 알 수 있다.

멀티모달 방식은 RNN의 학습 이후에 적용되기 때문에 RNN의 동작 특성으로 인해 발생하는 이미지 벡터의 크기 감소가 나타나지 않아 제일 많은 이미지 특징 벡터 크기를 유지한다. 이로 인해 멀티모달 레이어를 채택한 대부분의 모델들이 멀티모달 방식을 적용하고 있다. 하지만 멀티모달 방식은 RNN에 이미지 특징 벡터가 입력되지 않기 때문에 성능향상이 크지 않다. 따라서 본 연구에서는 Parallel Injection 방식과 함께 Multimodal 레이어의 이미지 특징 벡터를 최종 캡션 생성에서 사용함으로써 이미지 전체 특징을 고려한 캡션을 생성 할 수 있도록 하였다.

4.3 실험 결과

제안하는 이미지 캡션 모델의 타당성을 검증하기 위해 Google NIC[16], Hard Attention[17], m-RNN [18], Repeated Review[19]과 제안한 모델에 대하여 자연어 처리 분야에서 대표적으로 사용되는 Bi-Lingual Evaluation Understudy(BLEU)[25], Metric for Evaluation of Translation with Explicit ORdering (METEOR)[26] 기반 성능지표를 기준으로 평가하였으며, 실험 결과는 Table 1과 같이 나타난다. Flickr 8K 경우 BLEU-3에서 Repeated Review에 비해 점수가 낮지만 BLEU-4에서 다른 모델들에 비해 높은 점수를 가져 캡션 성능이 우수한 것을 객관적으로 확인 가능하다. 또한 Flickr 30K의 경우 전체적으로 제안하는 모델의 성능이 우수하며 MS-COCO의 경우에는 BLEU-1을 제외하면 다른 모델에 비해 점수가 낮지만 캡션 문장의 표현력을 구조적으로 분석하는 METEOR 점수를 통해 본 모델이 사람의 표현과 유사하게 표현되는 것을 알 수 있다. 실험 결과를 통해 구문의 존재를 판별하는 BLEU-2에서 4까지는 임베딩 레이어 Injection 방식을 적용하는 2개의 모델 모두 우수하게 나왔으나, 이미지 특징 벡터 소실로 인한 단어의 유무를 판단할 수 있는 BLEU-1 과 문맥적 특징을 분석하는 METEOR 점수는 제안한 Parallel Injection 방식이 기존의 다른 방식들에 비해 우수한 성능을 가지는 것을 나타낸다.

BLEU와 METEOR 점수의 결과를 통해 제안하는 모델은 Parallel Injection 기법을 통해 캡션 생성 과정에서 발생하는 이미지 특징 벡터의 소멸을 방지함으로써 이미지 자체가 가지는 정보를 기존의 모델에 비해 풍부하게 획득할 수 있으며, 일반적인 RNN이 아닌 Bi-GRU로 디코더를 구성함으로써 이전 및 이후 단어의 영향을 고려하여 캡션 생성 과정에서 전체 문맥에 맞춰 현재의 단어들을 수용하는 것을 알 수 있다. Fig. 3은 모델별 생성되는 자막의 샘플을 보여준다.

Fig. 3의 기준 캡션 대비 각 문장들의 특징을 통해 본 모델의 캡션 생성 특징을 분석할 수 있다. Fig. 3. (a)에서 Injection 방식을 적용하지 않은 NIC와 Hard attention은 보드를 타는 사람에 대해서는 표현

Table 1. BLEU and METEOR score for each model in the data set

Dataset	Model	BLEU				METEOR
		B1	B2	B3	B4	
Flickr 8K	Google NIC[16]	63.0	41.0	27.0	17.6	14.20
	Hard Attention[17]	67.0	45.7	31.4	21.3	20.30
	m-RNN[18]	58.0	28.0	23.0	14.2	-
	Repeated Review[19]	68.8	43.9	**31.6**	22.4	-
	Proposed model	**69.4**	**48.2**	30.7	**23.8**	**21.70**
Flickr 30K	Google NIC[16]	66.3	42.3	27.7	18.3	16.40
	Hard Attention[17]	66.9	43.9	29.6	19.9	18.46
	m-RNN[18]	60.0	41.0	28.0	19.0	-
	Repeated Review[19]	65.5	42.8	29.9	19.7	-
	Proposed model	**68.4**	**45.5**	**31.3**	**21.4**	**20.05**
MS-COCO	Google NIC[16]	66.6	46.1	32.9	24.6	23.70
	Hard Attention[17]	71.8	**50.4**	35.7	25.0	23.04
	m-RNN[18]	67.0	49.0	35.0	25.0	22.10
	Repeated Review[19]	72.7	49.6	**37.8**	**28.5**	24.30
	Proposed model	**73.5**	48.4	36.4	26.2	**24.74**

이 잘 진행되고 있으나, 옆에 있는 사람에 대해서는 캡션 생성과정에서 이미지 특징 벡터의 소멸로 인해 캡션에 등장하지 않는다. 특히 제안한 모델은 Parallel Injection 방식으로 인해 세밀한 특징이 남아있어 오히려 기준 캡션에 비해 벤치에 앉아있다는 표현과 같이 더 세밀한 묘사를 하는 것을 확인할 수 있다.

Fig. 3. (b)에서는 다른 3개의 모델에서는 여자 옷에 대한 색 표현이 캡션에 등장하지 않고, 현재 이미지의 환경에 대한 표현도 등장하지 않는다. 제안한 모델은 여성의 옷의 색과 현재 이미지의 환경이 도로에서 발생하는 것까지 정확하게 표현함으로써 세밀한 이미지 특징을 사용하는 것을 확인 할 수 있다.

Fig. 3. (c)에서는 NIC의 경우 단일객체에 대한 표현에 이미지 특징을 모두 소모하여 제일 처음에 나오는 자전거를 타는 사람의 헬멧과 같은 세부적인 특징은 캡션에 등장하지만 다른 사람들은 표현되지 않는다. 제안하는 모델은 자전거를 타는 사람과 쓰레기봉투 및 환경에 대한 표현까지 세밀한 캡션 생성이 가능한 것을 확인 할 수 있다.

마지막으로 Fig. 3. (d)는 4명의 어린이가 공중에 떠 있는 이미지로 NIC는 캡션 표현이 이미지의 내용과는 전혀 다른 문맥적 오류가 발생하며, Hard attention의 경우 학습데이터의 원반던지기와 관련된 이미지에서 사람 또는 강아지가 공중에 떠 있는 장면이 많아 원반던지기를 하는 상황이라는 캡션을 출력한다. 제안하는 모델은 정확하게 4명의 어린이가 공중에 점프하고 있는 상황을 캡션으로 출력함으로써 모델의 캡션 표현이 전체 이미지의 특징 및 특징 소멸에 강인함을 보여준다.

기준 캡션은 사람이 이미지를 보고 요약한 것으로 세밀한 표현은 생략하고 주요 객체 위주로 서술하는 경우가 많기 때문에 간결하다. 하지만 이러한 데이터 세트를 바탕으로 학습하더라도 세밀한 특징도 서술하는 캡션이 생성이 가능한 것을 확인할 수 있다.

5. 결 론

본 논문에서는 문장 표현력을 향상시키고 이미지 특징 벡터의 소멸을 방지할 수 있는 Parallel Injection 기법과 문맥에 맞는 문장의 순서를 생성하는 Bi-GRU을 적용한 디테일한 이미지 캡션 모델을 제안하였다. Parallel Injection 기법을 사용한 제안한 모델은 RNN의 재귀과정에서 이미지 특징 벡터의 소멸을 방지하기 위하여 이미지 특징 벡터를 단어 벡터와 융합하여 반복적으로 삽입해주어 문장 구성 요소 누락을 방지한다. 또한 양방향에서 획득하는 어휘 및 이미지 특징을 이용하는 Bi-GRU으로 디코더를 구성하여 문맥에 맞는 문장의 순서를 학습한다. 이 둘을 결합하여 기존의 Bi-RNN에 최적화 되어있지 않은 기존의 Injection 방식을 개선하여 이미지 및 문장

(a)

Reference : A boy in blue jumps his skateboard off some steps while his friends watch.
Google NIC : a man riding a skateboard over a hurdle.
Hard attention : a man riding a skateboard down a snow covered slope.
m-RNN : a man is sitting on a skateboard and people are sitting.
Proposed model : a man in blue is riding a skateboard and people are sitting on bench.

(b)

Reference : A blond woman in a blue shirt appears to wait for a ride.
Google NIC : a young boy wearing a red shirt is playing with a basket ball.
Hard attention : a person holding a cell phone in their hand.
m-RNN : a woman is standing outside holding a cell phone.
Proposed model : a woman in a blue shirt is holding a cell phone and is in the driveway.

(c)

Reference : A group of people ride bikes while holding onto large trash bags.
Google NIC : a man wearing a helmet rides a bike.
Hard attention : a group of people riding bikes down a street.
m-RNN : a group of people ride bikes on the road.
Proposed model : a group of people riding bike have trash bags on the road.

(d)

Reference : Four young kids have a picture taken of them while in midair.
Google NIC : a group of people are playing soccer.
Hard attention : a couple of young men playing a game of frisbee.
m-RNN : a group of people around playing outside.
Proposed model : a group of four children jumping in midair.

Fig. 3. Sample of the subtitle generated by the model. (a) A boy in blue, (b) A blond woman, (c) A group of people, and (d) Four young kids.

의 순서, 문장의 표현력을 모두 고려한 디테일한 캡션을 생성한다. 제안하는 모델은 BLEU와 METEOR 점수를 통해 모델의 성능을 객관적으로 비교하였고 기존의 캡션 모델에 비해 BLEU 점수는 최대 20.2점, METEOR 점수는 최대 3.65점이 향상되어 제안한 모델의 우수함을 보였다. 이 연구를 통하여 향후 영상의학 및 법영상 분야의 수준 높은 데이터세트가 구축되고 이를 학습한다면 해당 분야에 필요한 캡션을 생성함으로써 이미지의 자동 주석이나, 사용자의 질문에 대한 간단한 답변 표현에 적용이 가능 할 것으로 기대된다.

REFERENCE

[1] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional Neural Networks for Large-scale Remote-sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55, No. 2, pp. 645-657, 2017.

[2] D.H. Kim, J.E. Kim, J.H. Song, Y.J. Shin, and S.S. Hwang, "Image-based Intelligent Surveillance System Using Unmanned Aircraft," *Journal of Korea Multimedia Society*, Vol. 20, No. 3, pp. 437-445, 2017.

[3] S. Yu, S. Jia, and C. Xu, "Convolutional Neural Networks for Hyperspectral Image Classification," *Neurocomputing*, Vol. 219, pp. 88-98, 2017.

[4] P. Morales-Alvarez, A. Perez-Suay, R. Molina, and G. Camps-Valls, "Remote Sensing Image Classification with Large-scale Gaussian Processes," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 2, pp. 1103-1114, 2018.

[5] B. Gecer, G. Azzopardi, and N. Petkov, "Color-blob-based COSFIRE Filters for Object Recognition," *Image and Vision Computing*, Vol. 57, pp. 165-174, 2017.

[6] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFS," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 4, pp. 834-848, 2018.

[7] K. Grm, V. Stuc, A. Artiges, M. Caron, and H.K. Ekenel, "Strengths and Weaknesses of Deep Learning Models for Face Recognition Against Image Degradations," *The Institution of Engineering and Technology Biometrics*, Vol. 7, No. 1, pp. 81-89, 2017.

[8] J. Cleveland, D. Thakur, P. Dames, C. Phillips, T. Kientz, K. Daniilidis, et al., "Automated System for Semantic Object Labeling with Soft-object Recognition and Dynamic Programming Segmentation," *IEEE Transactions on Automation Science and Engineering*, Vol. 14, No. 2, pp. 820-833, 2017.

[9] X. Yang, W. Wu, K. Liu, P.W. Kim, A.K. Sangaiah, and G. Jeon, "Long-distance Object Recognition with Image Super Resolution: A Comparative Study," *IEEE Access*, Vol. 6, pp. 13429-13438, 2018.

[10] D. Marmanis, K. Schindler, J.D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an Edge: Improving Semantic Image Segmentation with Boundary Detection," *Journal of International Society for Photogrammetry and Remote Sensing*, Vol. 135, pp. 158-172, 2018.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv Preprint arXive:1409.1556*, 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelow, et al., "Going Deeper with Convolutions," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only Look once: Unified, Real-time Object Detection," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, pp. 91-99, 2015.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156-3164, 2015.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Proceeding of International Conference on Machine Learning*, pp. 2048-2057, 2015.

[18] 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:1412.6632*, 2014.

[19] J. Guan and E. Wang, "Repeated Review Based Image Captioning for Image Evidence Review," *Signal Processing: Image Communication*, Vol. 63, pp. 141-148, 2018.

[20] M. Schuster and K.K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681, 1997.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio,

"Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv Preprint arXiv:1412.3555*, 2014.

[22] 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.

[23] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions," *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 67-78, 2014.

[24] T.Y. Lin, M. Maire, S. Belongje, J. Hays, P. Perona, D. Ramanan, et al., "Microsoft Coco: Common Objects in Context," *arXiv Preprint arXiv:1405.0312*, 2014.

[25] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Asscociation for Computational Linguistics*, pp. 311-318, 2002.

[26] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *Proceeding of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65-72, 2005.

이 준 희

2016년 한국해양대학교 전자전기정보공학부 학사
2018년 한국해양대학교 전기전자공학과 석사
관심분야 : 컴퓨터 비전, 머신러닝, 자연어 처리

이 수 환

2016년 한국해양대학교 전자전기정보공학부 학사
2018년 한국해양대학교 전기전자공학과 석사
2018년~현재 한국해양대학교 전기전자공학과 박사 과정
관심분야 : 컴퓨터 비전, 머신러닝, 영상 처리

태 수 호

2018년 한국해양대학교 전자전기정보공학부 학사
2018년~현재 한국해양대학교 전기전자공학과 석사 과정
관심분야 : 머신러닝, 위치인식, 신호처리

서 동 환

1996년 경북대학교 전자공학부 학사
1999년 경북대학교 전자공학부 석사
2003년 경북대학교 전자공학부 박사
2004년~현재 한국해양대학교 전자전기정보공학부 교수
관심분야 : 신호 처리, 컴퓨터 비전, 센서 네트워크