KIPS Tr. Software and Data Eng.
Vol.6, No.4 pp.203~210 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.4.203

Design of a Deep Neural Network Model for Image Caption Generation

Dongha Kim† · Incheol Kim††

ABSTRACT

In this paper, we propose an effective neural network model for image caption generation and model transfer. This model is a kind of multi-modal recurrent neural network models. It consists of five distinct layers: a convolution neural network layer for extracting visual information from images, an embedding layer for converting each word into a low dimensional feature, a recurrent neural network layer for learning caption sentence structure, and a multi-modal layer for combining visual and language information. In this model, the recurrent neural network layer is constructed by LSTM units, which are well known to be effective for learning and transferring sequence patterns. Moreover, this model has a unique structure in which the output of the convolution neural network layer is linked not only to the input of the initial state of the recurrent neural network layer but also to the input of the multimodal layer, in order to make use of visual information extracted from the image at each recurrent step for generating the corresponding textual caption. Through various comparative experiments using open data sets such as Flickr8k, Flickr30k, and MSCOCO, we demonstrated the proposed multimodal recurrent neural network model has high performance in terms of caption accuracy and model transfer effect.

Keywords : Image Caption Generation, Deep Neural Network Model, Model Transfer, Multi-Modal Recurrent Neural Network

이미지 캡션 생성을 위한 심층 신경망 모델의 설계

김 동 하† · 김 인 철††

요 약

본 논문에서는 이미지 캡션 생성과 모델 전이에 효과적인 심층 신경망 모델을 제시한다. 본 모델은 멀티 모달 순환 신경망 모델의 하나로서, 이미지로부터 시각 정보를 추출하는 컨볼루션 신경망 층, 각 단어를 저차원의 특징으로 변환하는 임베딩 층, 캡션 문장 구조를 학습하는 순환 신경망 층, 시각 정보와 언어 정보를 결합하는 멀티 모달 층 등 총 5 개의 계층들로 구성된다. 특히 본 모델에서는 시퀀스 패턴 학습과 모델 전이에 우수한 LSTM 유닛을 이용하여 순환 신경망 층을 구성하며, 캡션 문장 생성을 위한 매 순환 단계마다 이미지의 시각 정보를 이용할 수 있도록 컨볼루션 신경망 층의 출력을 순환 신경망 층의 초기 상태뿐만 아니라 멀티 모달 층의 입력에도 연결하는 구조를 가진다. Flickr8k, Flickr30k, MSCOCO 등의 공개 데이터 집합들을 이용한 다양한 비교 실험들을 통해, 캡션의 정확도와 모델 전이의 효과 면에서 본 논문에서 제시한 멀티 모달 순환 신경망 모델의 높은 성능을 확인할 수 있었다.

키워드 : 이미지 캡션 생성, 심층 신경망 모델, 모델 전이, 멀티 모달 순환 신경망

1. 서 론

이미지(image)로부터 그 이미지가 어떤 내용(content)을 담고 있는가를 표현하는 문장(sentence)들을 자동으로 생성하는 기술을 이미지 캡션 생성(image caption generation) 기술이라고 한다[1, 2]. 예컨대, Fig. 1에는 이미지 캡션 예들을 보여주고 있는데, 위쪽에는 이미지들이 주어져 있고, 아래쪽에는 각 이미지에 담긴 내용을 설명하는 캡션 문장들이 주어져 있다. 이와 같이 이미지와 캡션 문장들이 훈련 데이터로 주어지면, 이들을 토대로 이미지의 시각 정보(visual information)와 캡션 문장의 언어 정보(language information) 간의 관계를 스스로 학습하여 새로운 이미지에 대한 캡션을 자동으로 생성해내는 기술을 이미지 캡션 생성 기술이라 부른다. 이미지 캡션 생성 기술은 시각 인식 기술과 자연어 처리 기술이 함께 요구되기 때문에 매우

※ 본 연구는 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구 과제임(No. 10060086, 개인 서비스용 로봇을 위한 지능-지식 집약·개방·진화형 로봇지능 소프트웨어 프레임워크 기술 개발).
※ 이 논문은 2016년도 한국정보처리학회 추계학술발표대회에서 '이미지 캡션 생성을 위한 심층 신경망 모델 학습과 전이'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터과학과 교수
Manuscript Received : December 15, 2016
Accepted : December 28, 2016
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

복잡하고 어려운 기술이다. 하지만 이 기술은 이미지 검색(image retrieval), 유아 교육(early childhood education), 시각 장애인들을 위한 길 안내(navigation for the blind)와 같은 다양한 응용 분야들에 유용하게 활용될 수 있는 중요한 기술이다[3, 4].

최근 영상 인식 분야에서 물체 인식과 탐지 등에 컨볼루션 신경망(CNN, Convolution Neural Network)이 활발히 이용되고 있으며, 또한 자연어 처리 분야에서도 기계 번역 등에 순환 신경망(RNN, Recurrent Neural Network)의 활용이 큰 성공을 보였다. 이와 같은 성공 사례들에 힘입어, 최근에는 이미지 캡션 생성에도 심층 신경망들을 활용해보려는 노력들이 활발해졌다. 특히 자연어 기계 번역을 위한 시퀀스 패턴 학습에 큰 효과를 보았던 순환 신경망(RNN)은 이미지를 표현하는 자연어 캡션 문장 생성에도 큰 도움을 주는 것으로 알려져 있다.

Fig. 1. Image Caption Examples

최근 연구들을 통해 제시된 이미지 캡션 생성을 위한 다양한 순환 신경망 모델들 중에서 현재 가장 보편적인 모델은 멀티 모달 순환 신경망(multimodal recurrent neural network) 모델로서, 크게 언어 모델 부분(language model part)과 시각 모델 부분(visual model part), 그리고 이들을 결합하는 멀티 모달 부분(multimodal part)들로 구성된다. 하지만 이미지 캡션 생성을 위한 멀티 모달 순환 신경망 모델에 관한 몇 가지 중요한 질문들은 아직 명확히 해결되지 않은 상태로 남아 있다. 그중 첫 번째 질문은 시각 모델과 언어 모델의 결합 방식에 관한 것으로서, 이미지의 시각 정보를 추출하는 컨볼루션 신경망(CNN)의 출력을 캡션 문장 생성을 위한 순환 신경망(RNN)에 어떤 방식으로 연결할 것인가이다. 기존 연구들에서는 이미지에서 추출한 시각 정보들을 순환 신경망의 입력으로 한번만 사용하는 방식과 이들을 캡션 문장 생성을 위한 매 단계에서 이용할 수 있도록 멀티 모달 층(multimodal layer)에도 연결하는 두 가지 방식이 시도되었다. 그동안 서로 엇갈린 실험 결과들이 보고된 적은 있지만, 어느 연결 방식이 더 우수한 방식인지 명확히 밝혀진 바는 아직 없다.

이미지 캡션 생성을 위한 멀티 모달 순환 신경망 모델에 관한 두 번째 질문은 순환 신경망 층(RNN layer)을 어떤 유닛(unit)들로 구성해야 하는가이다. 그동안 심층 신경망 연구자들은 순환 신경망(RNN)의 깊은 구조로 인한 가중치 소멸 문제(vanishing gradient problem)를 극복하기 위해, LSTM (Long Term Short Memory)[6], GRU (Gated Recurrent Unit)[7] 등과 같은 새로운 순환 신경망 유닛들을 개발하였다. GRU는 LSTM에 비해 훨씬 적은 수의 내부 게이트(gate)들을 포함함으로써, LSTM에 비해 학습 시간을 단축할 수 있는 장점이 있는 것으로 알려져 있다[5]. 하지만, 서로 다른 이 두 가지 유형의 순환 신경망 유닛들이 이미지 캡션 성능 면에서 어떤 것이 더 우수한지 명확히 비교된 사례는 없다. 또한, 하나의 영역(domain)에서 학습한 순환 신경망 모델을 다른 영역들에서 이미지 캡션 생성을 위해 활용하고자 할 때, 즉 영역들 간의 모델 전이(model transfer)가 필요할 때, 과연 어떤 순환 신경망 모델과 유닛들이 더 유리한지에 대해 구체적으로 연구한 결과는 아직 없는 것으로 알고 있다.

본 논문에서는 앞서 언급한 질문들에 답하기 위해, 효과적인 이미지 캡션 생성을 위한 멀티 모달 순환 신경망 모델을 제시한다. 본 모델에서는 시퀀스 패턴 학습과 모델 전이에 우수한 LSTM 유닛들로 순환 신경망 층(RNN layer)을 구성하며, 컨볼루션 신경망 층(CNN layer)을 통해 추출되는 시각 정보들을 매번 다음 단계 캡션 단어를 예측하는데 이용할 수 있도록 순환 신경망 층(RNN layer)의 초기 상태뿐만 아니라 멀티 모달 층(multimodal layer)의 입력에도 연결하는 구조를 가진다. Flickr8k, Flickr30k, MSCOCO 등 서로 다른 공개 데이터 집합들을 이용한 다양한 비교 실험을 통해, 본 논문에서 제시한 멀티 모달 순환 신경망 모델의 우수성을 입증한다.

2. 관련 연구

이미지 캡션 생성 모델에 대한 선행 연구들은 본 논문에서 다루고자 하는 구조적 관점에서 크게 단순 신경망을 이용한 연구들과 멀티 모달 순환 신경망 이용한 연구들로 나누어 볼 수 있다. 또한 순환 신경망 층을 어떤 유닛으로 구현했는가도 구분해 볼 수 있다.

먼저 Vinyals의 연구[2]에서는 기계 번역에 효과적으로 사용된 한 쌍의 인코더 순환 신경망(encoder RNN)과 디코더 순환 신경망(decoder RNN) 구성에 영감을 얻어, 이미지 캡션 생성을 위한 새로운 심층 신경망 모델을 제시하였다. 이 모델에서는 인코더 순환 신경망 대신, 주로 영상 분류와 물체 인식 등에 적용되어 오던 컨볼루션 신경망(CNN)을 이미지 캡션 생성을 위한 이미지 인코더(image encoder)로 사용하는 방식을 채택하였다. 그리고 이 디코더 순환 신경망을 LSTM 유닛들로 구성하였으며, 컨볼루션 신경망(CNN)을 통해 추출된 이미지의 시각 정보를 이용하였다. 하지만 Vinyals의 연구에서는 디코더 순환 신경망(decoder RNN)의 첫 단계 입력으로만 제공하기 때문에 시각 정보를 적극적으로 이용했다고 할 수 없다. 한편, Vinyals의 모델을 확장한 Xu의 연구[3]에서는 이미지에서 주목할 중요한 관심 영역들(attention)을 먼저 찾아내고, 이들을 토대로 이미지 캡션을 생성하는 순환 신경망 모델을 제안하였다. 이 모델에서도 이미지로부터 시각 특징을 추출하기 위해서는 컨볼루션 신

경망 층(CNN layer)을 이용하며, 순환 신경망 층(RNN layer)은 LSTM 유닛들로 구성하였다. 두 연구 모두 효과적인 캡션 생성 모델에 대한 비교 실험은 수행하지 않았다는 한계점이 있다.

Mao의 연구[4]에서는 보다 언어 모델을 강화하기 위한 멀티 모달 순환 신경망(multimodal RNN) 모델을 제시하였다. 이 모델에서는 시각 모델 학습을 위한 컨볼루션 신경망 층(CNN layer)외에 언어 모델 학습을 위한 두 개의 임베딩 층(embedding layer)과 하나의 순환 신경망 층(RNN layer)을 두었고, 언어 모델과 시각 모델의 결합을 위한 별도의 멀티 모달 층(multimodal layer)을 두었다. 이 모델에서 순환 신경망 층은 확장형 단순 순환 신경망 유닛들로 구성하였다. Lee의 연구[5]에서는 Mao의 모델에 언어 모델 부분의 연결 구조를 다양화한 확장형 멀티 모달 순환 신경망 모델을 제안하였다. 이 모델에서는 순환 신경망 층(RNN layer)의 구성을 위해 GRU 유닛들을 이용하였고, 컨볼루션 신경망 층(CNN layer)의 출력인 이미지 시각 정보는 멀티 모달 층(multimodal layer)에 공급하는 연결 구조를 사용하였다. Mao와 Lee의 연구에서는 캡션 생성 모델에 대한 비교 실험은 진행하였으나 순환 신경망 유닛에 대한 비교 실험은 수행되지 않았다.

Lisa의 연구[1]에서는 학습 데이터에서 등장하지 않았던 새로운 물체가 포함된 이미지가 입력으로 들어왔을 때, 학습 데이터 내의 물체가 아닌 새로운 물체가 포함된 캡션 생성을 위한 방법을 제시하였다. 학습된 물체 이미지 간의 가중치의 전이(transfer)를 통해 새로운 캡션을 생성하였다. Lisa의 연구에서도 멀티 모달 순환 신경망 모델을 사용하였다. 시각 정보는 미리 훈련된 CNN 모델을 이용하여 추출하였고, 순환 신경망 유닛으로는 LSTM 유닛을 사용하였다. 시각 정보와 언어 정보는 멀티 모달 층에서 결합하는 방법을 사용하였다. 하지만 Lisa의 연구는 전이에 효과적인 순환 신경망 모델에 대한 비교 실험을 수행하지 않았다는 한계점이 있다.

3. 이미지 캡션 생성

본 논문에서는 Fig. 2와 같이 이미지 캡션 생성을 위해 크게 두 가지 유형의 신경망들을 포함한 이미지 캡션 생성 모델을 사용한다. 그 중 하나는 이미지의 시각 모델을 학습하는 컨볼루션 신경망(CNN)이고, 다른 하나는 캡션의 언어 모델을 학습하는 순환 신경망(RNN)이다.

좀 더 구체적인 심층 신경망 모델은 이미지로부터 시각 특징을 추출하는 컨볼루션 신경망 층(CNN layer), 각 단어를 저차원의 특징으로 변환하는 임베딩 층(embedding layer), 캡션 문장 구조를 학습하는 순환 신경망 층(RNN layer), 시각 특징과 언어 특징을 결합하는 멀티 모달 층(multimodal layer) 등 총 5 개의 계층(layer)들로 구성된 멀티 모달 순환 신경망 모델이다. 이러한 이미지 캡션 생성을 위한 멀티 모달 순환 신경망 모델의 중요한 설계 요소들로 (1) 컨볼루션 신경망 층(CNN layer)의 시각 정보 출력을 캡션 언어 정보를 학습하는 순환 신경망 층(RNN layer)과 연결하는 연결 구조(connection structure)와 (2) 순환 신경망 층(RNN layer)을 구성하는 유닛들의 종류(type of RNN units)를 결정하는 일 등이다. 이들은 모델 학습 시간(model learning time)과 캡션 정확도(caption accuracy), 모델 전이 효과(model transfer effect) 등 다양한 면에서 성능에 큰 영향을 미친다. 따라서 이러한 점들을 종합적으로 고려하여, 이러한 설계 요소들을 신중히 결정해야 한다.

3.1 시각 정보 연결 구조

본 논문에서는 제안하는 이미지 캡션 자동 생성을 위한 멀티 모달 순환 신경망 모델은 Fig. 3과 같이 서로 다른 시각 정보 연결 구조를 가질 수 있다. Fig. 3의 (a)는 컨볼루션 신경망 층(CNN layer)의 시각 정보 출력을 순환 신경망 층의 초기 값으로 이용하는 구조를 나타낸다. 이러한 시각 정보 연결 구조를 사용할 경우, 이미지의 시각 정보는 캡션 생성을 위한 순환 신경망의 첫 단계에 한번만 사용된다. 반면에, Fig. 3의 (b)는 컨볼루션 신경망 층(CNN layer)의 시각 정보 출력을 멀티 모달 층(multimodal layer)에 연결하는 구조를 나타낸다. 이러한 시각 정보 연결 구조를 사용할 경우, 멀티 모달 층에서 시각 정보와 언어 정보를 결합하여 순환 신경망의 매 단계마다 사용한다. Fig. 3(b) 구조의 순환 신경망의 초기 값은 0으로 설정하여 훈련하였다. 마지막으로 Fig. 3의 (c)는 시각 정보의 출력을 순환 신경망 층의 초기 값으로도 사용하고, 멀티 모달 층에도 연결하였다. 멀티 모달 층에서 시각 정보를 캡션 단어를 생성하는 매 단계마다 사용하는 Fig. 3(b)의 연결 구조는 (a)의 연결 구조에 비해 모델 학습과 캡션 생성에 소요되는 시간은 증가할 수 있으나, 캡션의 정확도는 개선될 수 있을 것으로 판단한다. 또한 Fig. 3(a)의 연결 구조와 (b)의 연결 구조에서의 시각 정보 활용을 모두 이용하는 (c)의 연결 구조의 정확도는 더

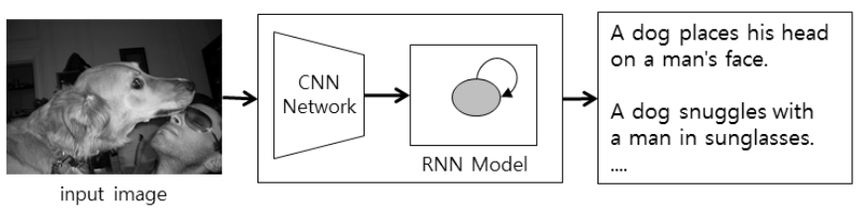

Fig. 2. Concept Model for Image Caption Generation

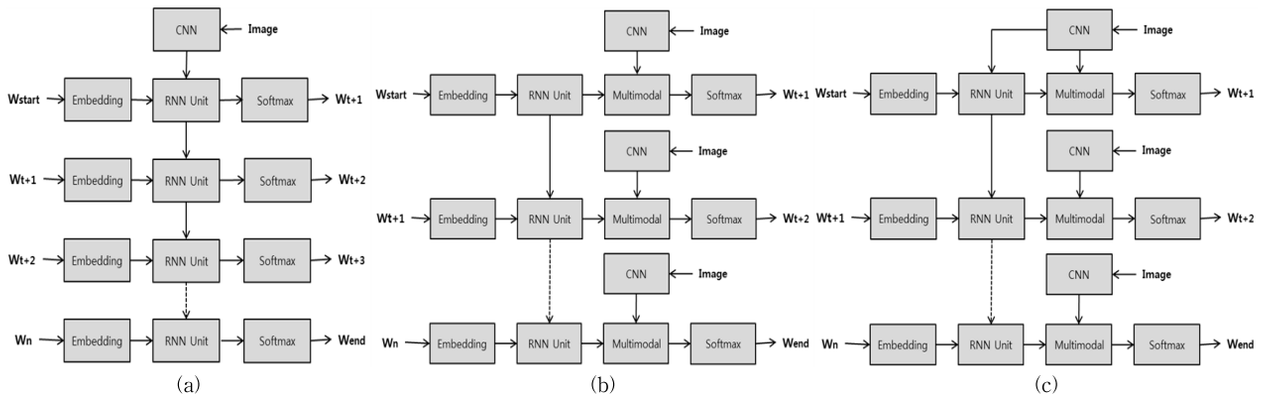

Fig. 3. Visual Information Connection Architecture

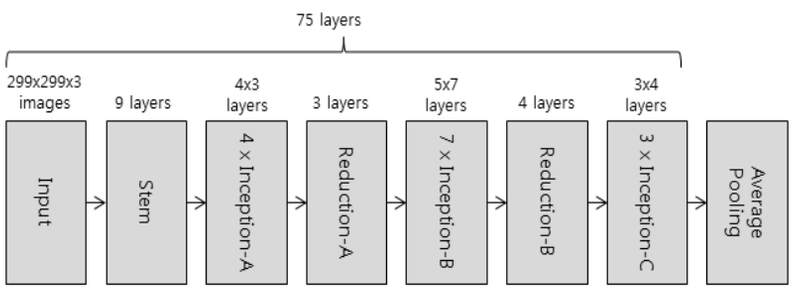

Fig. 4. Architecture of Inception v4

욱 개선될 것이라 판단된다. 따라서 본 논문에서는 Fig. 3(c)와 같은 연결 구조를 갖는 멀티 모달 순환 신경망 모델을 사용한다. 그리고 이미지로부터 시각 정보를 추출하기 위해서 Inception v4 컨볼루션 신경망(CNN)[8]을 이용하였다. Fig. 4는 Inception v4 컨볼루션 신경망의 구조이다. Inception v4 컨볼루션 신경망은 계층 수를 이전 연구들보다 증가시키고, 계층들을 결합(merge)하는 Inception 구조를 통해 학습 속도와 성능을 크게 개선시킨 신경망이다. Inception v4 컨볼루션 신경망은 299×299×3 크기의 이미지를 입력으로 받으며, 8개의 레이어로 구성된 줄기(Stem) 단계, 12개, 35개, 12개 레이어로 구성된 3개의 인셉션(Inception) 단계, 3개, 4개 레이어로 구성된 2개의 축소(Reduction) 단계로 구성되어 있다. 총 75개의 레이어를 가지고, 2048 차원의 특징을 생성한다.

3.2 순환 신경망 유닛

가중치 소멸 문제(vanishing gradient problem)을 극복하기 위해 새로 개발된 대표적인 순환 신경망 유닛들로는 LSTM과 GRU 등이 있다. Fig. 5의 (a)와 (b)는 각각 LSTM 유닛과 GRU 유닛의 내부 구조를 나타낸다. 하나의 LSTM 유닛은 Fig. 5의 (a)와 같이 입력 게이트(input gate), 출력 게이트(output gate), 망각 게이트(forget gate) 등 총 세 개의 게이트들로 셀 갱신(cell update)과 출력(output) 제어가 가능한 하나의 메모리 셀(memory cell)을 나타낸다. LSTM 내부 파라미터(parameters)를 결정하는 수식은 아래와 같다.

$$i_t = \sigma(x_t U^i + s_{t-1} W^i) \tag{1}$$

$$f_t = \sigma(x_t U^f + s_{t-1} W^f) \tag{2}$$

$$o_t = \sigma(x_t U^o + s_{t-1} W^o) \tag{3}$$

$$g = \tanh(x_t U^g + s_{t-1} W^g) \tag{4}$$

$$c_t = c_{t-1} \circ f + g \circ i \tag{5}$$

$$s_t = \tanh(c_t) \circ o \tag{6}$$

Equation (1)~(6)과 같이 LSTM 유닛은 게이트들을 포함해 많은 수의 내부 파라미터들을 포함하고 있어서, 많은 훈련 데이터와 긴 학습 시간을 요구한다. 하지만 다음 시퀀스를 결정하는데 많은 가중치가 사용되기 때문에 비교적 정확한 캡션 생성 모델을 얻을 수 있다. 반면에 하나의 GRU 유닛은 Fig. 5의 (b)와 같이 갱신 게이트(update gate), 리셋 게이트(reset gate) 등 단 두 개의 게이트로 셀 갱신과 출력을 조절할 수 있는 순환 신경망 유닛이다. GRU 유닛의 내부 파라미터(parameters)를 결정하는 수식은 아래와 같다.

$$u_t = \sigma(x_t U^u + s_{t-1} W^u) \tag{7}$$

$$r_t = \sigma(x_t U^r + s_{t-1} W^r) \tag{8}$$

$$h = \tanh(x_t U^h + (s_{t-1} \circ r) W^h) \tag{9}$$

$$s_t = (1 - z) \circ h + z \circ s_{t-1} \tag{10}$$

Equation (7)~(10)과 같이 GRU 유닛은 LSTM 유닛에 비해 학습해야 될 내부 파라미터들의 수가 적기 때문에, 상대적으로 짧은 모델 학습 시간을 요구한다. 하지만 비교적 단순한 내부 구조로 인해 캡션 생성 모델의 정확도는 LSTM 유닛에 비해 낮을 것으로 예상한다. 따라서 본 논문의 멀티모달 순환 신경망 모델에서는 학습 시간 면에서 약간 유리한 GRU 유닛 대신 캡션 생성 모델의 정확도가 높은 LSTM 유닛을 선택하였다.

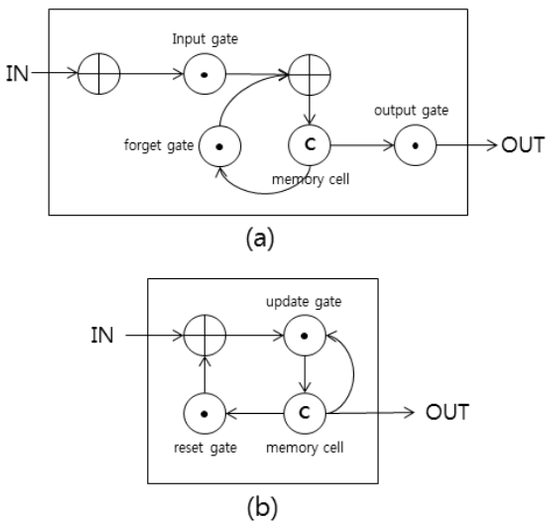

Fig. 5. Architecture of (a) LSTM and (b) GRU

3.3 모델 전이

새로운 영역(domain)에서 이미지 캡션 생성 작업을 위해 매번 그 영역에서 수집한 대규모 훈련 데이터 집합으로 신규 모델을 학습하는 것은 매우 낭비적인 방식이다. 일반적으로 하나의 영역에서 학습한 모델이나 지식을 다른 영역들에서 효과적으로 재활용하는 기술을 모델 전이(model transfer)라고 부른다. 이미 학습해둔 이미지 캡션 생성용 순환 신경망 모델을 다른 영역들에서 이미지 캡션 생성을 위해 재활용하고자 할 때도 모델 전이(model transfer)가 필요하다. 이러한 모델 전이의 효과를 고려한다면 과연 어떤 순환 신경망 모델을 이용하는 것이 유리한가하는 판단을 순환 신경망 설계에 반영할 수 있다. 본 논문에서는 앞서 설명한 두 가지 순환 신경망 유닛들 중에서 다양한 조절 게이트들을 포함한 LSTM 유닛이 비교적 단순한 GRU 유닛에 비해 모델 전이에도 더 효과적이라고 판단하였다.

이러한 가설을 입증하기 위해, 본 논문에서는 Fig. 6과 같이 원래 도메인(source domain)과 목표 도메인(target domain)의 다양한 변화에 대한 모델 전이 실험을 수행한다.

4. 실험 및 평가

본 논문에서는 성능 실험을 위해 Flick8k, Flickr30k, MSCOCO[10] 등 세 개의 공개 데이터 집합을 사용하였다. Flickr8k과 Flickr30k는 Flickr에서 추출한 8,000개, 30,000개의 이미지와 캡션들로 구성되어 있으며, MSCOCO는 국제경진대회용으로 수집한 대규모 이미지 캡션 데이터 집합이다. 실험을 위한 심층 신경망 모델 학습을 위해서 Python 딥러닝 라이브러리인 TensorFlow를 이용하였으며, 실험은 Ubuntu 14.04 LTS 64bit 컴퓨터 환경에서 수행되었다. 훈련 및 검증 데이터와 테스트 데이터의 분포는 Flickr8k의 경우 훈련 데이터 6,000개, 검증 데이터 1000개, 테스트 데이터 1000개를 사용하였다. Flickr30k의 경우는 훈련 데이터 25,381개, 테스트 데이터 3,000개, 나머지 데이터는 검증에 사용하였다. MSCCCO의 경우는 훈련 데이터 82,783개, 테스트 데이터 40775개를 사용하였다. 각 이미지에는 다섯 문장 이상의 캡션이 함께 제공된다.

Table 1. Caption Accuracy on Flickr8k

Dataset	type	BLEU_1	BLEU_2	BLEU_3	BLEU_4
Flickr8k	(a) - GRU	0.564	0.378	0.245	0.156
	(b) - GRU	0.589	0.404	0.269	0.172
	(c) - GRU	0.592	0,410	0.274	0.178
	(a) - LSTM	0.582	0.395	0.261	0.168
	(b) - LSTM	0.595	0.408	0.273	0.178
	(c) - LSTM	**0.615**	**0.419**	**0.287**	**0.192**

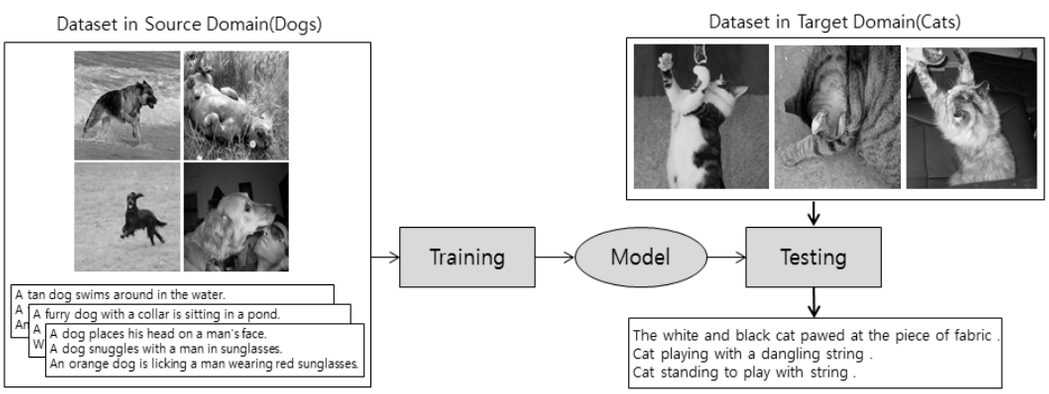

Fig. 6. Model Transfer

Table 2. Caption Accuracy on Flickr30k

Dataset	type	BLEU_1	BLEU_2	BLEU_3	BLEU_4
Flickr30k	(a) - GRU	0.587	0.392	0.257	0.168
	(b) - GRU	0.612	0.423	0.286	0.192
	(c) - GRU	0.617	0.421	0.280	0.187
	(a) - LSTM	0.604	0.410	0.274	0.184
	(b) - LSTM	0.619	0.426	0.287	0.193
	(c) - LSTM	**0.625**	**0.432**	**0.289**	**0.194**

Table 3. Caption Accuracy on MSCOCO

Dataset	type	BLEU_1	BLEU_2	BLEU_3	BLEU_4
MSCOCO	(a) - GRU	0.679	0.502	0.367	0.272
	(b) - GRU	0.684	0.502	0.364	0.266
	(c) - GRU	0.691	0.505	0.364	0.265
	(a) - LSTM	0.674	0.493	0.356	0.259
	(b) - LSTM	0.683	0.502	0.369	0.274
	(c) - LSTM	**0.692**	**0.517**	**0.381**	**0.283**

첫 번째 실험에서는 Fig. 3의 (a)와 (b)에 제시된 서로 다른 시각 정보 연결 구조와 LSTM, GRU 등 서로 다른 순환 신경망 유닛에 따른 캡션의 정확도와 모델 학습시간을 비교해 보았다. Table 1, Table 2, Table 3은 시각 정보 연결 구조 {(a), (b), (c)}와 유닛 종류 {GRU, LSTM}의 서로 다른 여섯 가지 조합들에 따른 캡션 정확도(caption accuracy)를 평가한 실험 결과를 나타낸다. 캡션 정확도는 Equation (11)과 (12)에 정의된 N 그램 문장 단위 평가 척도인 BLEU-N 계산식[9]을 이용하여 평가하였다. 식에서 r은 정답인 문장 수를, c는 생성된 문장 수를 나타낸다.

$$BP = \min(1, e^{1-\frac{r}{c}}) \tag{11}$$

$$BLEU-N = BP \cdot e^{\frac{1}{N}\sum_{n=1}^{N}\log(p_n)} \tag{12}$$

실험 결과, 본 논문에서 제안한 (c) 연결 구조와 LSTM 유닛의 조합((c)-LSTM)이 다른 모든 연결 구조와 유닛의 조합들에 비해 Flickr8k, Flickr30k, MSCOCO 등 거의 모든 데이터 집합들에서 공통적으로 가장 높은 캡션 정확도를 보여주었다. 또한, (a), (b) 연결 구조에 비해 (c) 연결 구조의 캡션 정확도 증가는 모든 데이터 집합들에서 매우 분명하며, GRU 유닛에 비해 LSTM 유닛의 캡션 정확도 증가도 MSCOCO 데이터 집합의 일부를 제외한 Flickr8k, Flickr30k 등에서는 분명히 확인할 수 있다.

Table 4는 시각 정보 연결 구조 {(a), (b), (c)}와 유닛 종류 {GRU, LSTM}의 서로 다른 여섯 가지 조합들에 따른 모델 학습 시간들을 비교 실험한 결과들을 나타낸다. 본 실험에서 모델 학습 시간은 각 모델의 에러 함수 값이 2.3 이하로 감소할 때까지 학습에 소요된 시간을 측정하였다. 실험 결과, (a) 연결 구조와 GRU 유닛의 조합이 가장 짧은 학습 시간을 소모하였다. 연결 구조만으로 비교했을 때는 데이터양이 적은 flickr8k를 제외하면 (a), (b), (c) 순으로 학습 시간이 짧았다. 순환 신경망 유닛만 비교해보면, 예상한대로 전반적으로 LSTM 유닛의 학습 시간이 GRU의 경우에 비해 좀 더 긴 것을 확인할 수 있다.

Table 4. Model Learning Time

	flickr8k	flickr30k	MSCOCO
(a)+GRU	25m	1h 35m	11h 26m
(b)+GRU	28m	1h 36m	11h 50m
(c)+GRU	27m	1h 36m	11h 52m
(a)+LSTM	28m	1h 40m	14h 55m
(b)+LSTM	31m	1h 46m	14h 41m
(c)+LSTM	29m	1h 49m	14h 49m

Table 5. Model Transfer with GRU

Training \ Test	Flickr8k	Flickr30k	MSCOCO
Flickr8k		0.548	0.460
Flickr30k	0.636		0.514
MSCOCO	0.551	0.556	

Table 6. Model Transfer with LSTM

Training \ Test	Flickr8k	Flickr30k	MSCOCO
Flickr8k		0.565	0.470
Flickr30k	0.614		0.500
MSCOCO	0.552	0.560	

두 번째 실험에서는 Flick8k, Flickr30k, MSCOCO 등 서로 다른 데이터 집합들을 이용하여, LSTM 유닛과 GRU 유닛을 채용한 멀티 모달 순환 신경망 모델들 간의 모델 전이 효과를 분석해보았다. Table 5와 Table 6은 모델 학습을 위한 훈련 데이터 집합과 캡션 생성을 위한 테스트 데이터 집합의 서로 다른 조합들에 대해, 각각 GRU 유닛과 LSTM 유닛의 모델 전이 실험 결과를 나타낸다. 전이 실험에는 시각 정보 연결 구조 (c)를 이용하였다. 두 표에 제시된 모델 전이 결과는 BLEU_1로 측정한 캡션 정확도이다. 실험 결과에서, Flickr30k를 훈련 데이터 집합으로 실험한 경우를 제외하면, 모든 실험 조합들에서 본 논문의 LSTM 유닛을 사용한 멀티 모달 순환 신경망 모델이 GRU 유닛을 사용한 모델에 비해 모델 전이 결과로 더 높은 캡션 정확도를 얻었음을 알 수 있다. 이러한 실험 결과들은 생성되는 캡션의 정확도와 모델 전이의 효과 면에서 본 논문에서 제안한 멀티 모달 순환 신경망 모델의 시각 정보 연결 구조와 LSTM 순환 신경망 유닛의 우수성을 확인해주는 결과로 볼 수 있다.

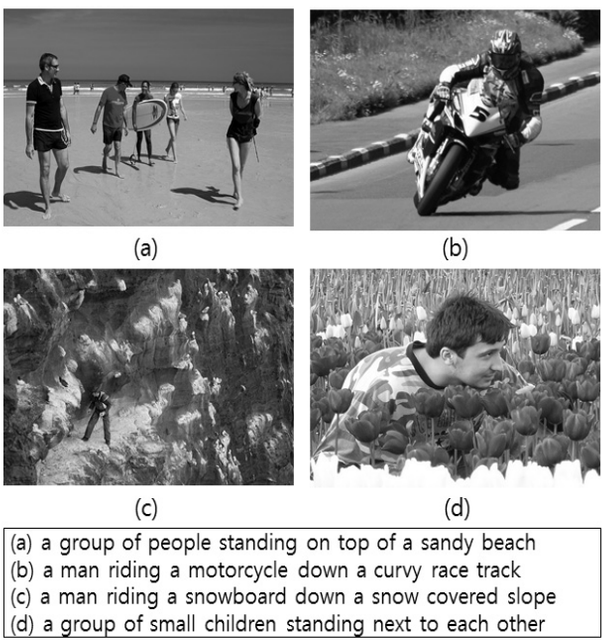

Fig. 7. Some Results of Caption Generation

Fig. 7은 새로운 이미지를 Flickr30k 데이터로 학습된 모델에 적용하여 캡션 생성을 수행한 결과이다. Fig. 7에서 (a)와 (b)는 해변 위에 서 있는 사람과 오토바이를 타고 있는 남자를 비교적 잘 표현하였다. 이에 반해 (c)와 (d)는 잘못된 캡션이 생성되었는데, (c)와 같은 경우에는 눈 내린 산에서 남자가 등산을 하고 있지만, 생성된 캡션은 남자가 눈 내린 산에서 스노우 보드를 타는 것으로 표현되었다. 이는 눈 내린 산과 스노우 보드가 함께 등장하는 훈련 데이터가 많기 때문이다. 그리고 (d)와 같은 경우는 이미지와 관련성이 적은 잘못된 캡션이 생성되었는데, 이는 다양한 꽃에 대한 학습 데이터가 부족하고, 중요한 물체인 아이가 비교적 두각 되지 않아서 캡션 생성에 오류가 발생한 것이다. 이러한 오류를 줄이기 위해서 캡션 생성 모델을 생성할 때 더 다양한 사물이 등장하는 학습 데이터를 사용해야 할 것이다.

5. 결 론

본 논문에서는 이미지 캡션 생성에 효과적인 심층 신경망 모델을 제시하였다. 본 모델은 멀티 모달 순환 신경망 모델의 하나로서, 순환 신경망 층은 시퀀스 패턴 학습과 모델 전이에 우수한 LSTM 유닛들로 구성되며, 시각 정보를 제공하는 컨볼루션 신경망 층의 출력은 순환 신경망 층의 초기 상태뿐만 아니라 캡션 문장 생성을 위한 매 순환 단계마다 멀티 모달 층의 입력으로 공급되는 연결되는 구조를 가진다. Flickr8k, Flickr30k, MSCOCO 등의 공개 데이터 집합들을 이용한 비교 실험을 통해, 본 논문에서 제안한 멀티 모달 순환 신경망 모델의 우수성을 확인할 수 있었다. 최근에는 효과적인 캡션 생성을 위해 영상 안에 주목할 만한 특정 영역들에 대한 주의 집중(attention) 기술에 관한 연구들이 활발하다. 이러한 주의 집중 기술을 채용한 이미지 캡션 생성 시스템들이 Flicker8K, Flicker30K, MS COCO 등의 데이터 집합들에서 BLEU_1 정확도가 각각 0.68, 0.66, 0.71을 상회함으로써, 현재는 최고 성능을 나타내고 있다. 주의 집중 기술을 채용하지 않은 본 논문의 방법은 아직 이러한 시스템들에 비해서는 조금 못미치는 성능을 나타내고 있다. 따라서 계획하고 있는 향후 연구로는 이미지의 특정 영역과 그것에 대응하는 캡션 문장 요소들을 연결할 수 있는 효과적인 주의 집중 기술을 개발하는 것, 이미지로부터 추출한 시각 정보와 캡션 문장으로부터 추출한 언어 정보를 좀 더 효과적으로 결합할 수 있는 멀티 모달 층을 개발하는 것 등이 있다.

References

[1] Lisa Anne Hendricks et al., “Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data,” *Proc. of IEEE Conf. on CVPR*, 2016.

[2] Oriol Vinyals and Alexander Toshev et al., “Show and Tell: A Neural Image Caption Generator,” *Proc. of the IEEE, Conf. on CVPR*, 2015.

[3] Kevin Xu and Jimmy Lei Ba et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” *Proc. of. ICML*. 2015.

[4] Junhua Mao, Wei Xu, and Yi Yang et al., “Deep Captioning with Multimodal Recurrent Neural Networks (M-RNN),” *Proc. of. ICLR*, 2015.

[5] Changki Lee, “Image Caption Generation using Recurrent Neural Network,” *Journal of KIISE*, Vol.43, No.8, pp.878-882, 2016.

[6] Hochreiter, Sepp, and Jürgen Schmidhuber, “Long Short-Term Memory,” *Neural Computation,* Vol.9, No.8, pp.1735-1780, 1997.

[7] Chung, Junyoung et al., “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” arXiv preprint arXiv:1412.3555, 2014.

[8] Szegedy, Christian, Sergey Ioffe et al., “Inception-v4, Inception-Resnet and The Impact of Residual Connections on Learning,” arXiv preprint arXiv:1602.07261, 2016.

[9] Papineni Kishore, Rouskos Salim et al., “BLEU: a Method for Automatic Evaluation of Machine Translation,” *Proc. of ACL*, pp.311-318, 2002.

[10] Lin Tsung-Yi and Maire Michael et al., “Microsoft COCO: Common Objects in Context,” *Proc. of ECCV*, Springer International Publishing, 2014.

김 동 하

e-mail : kdh2040@kyonggi.ac.kr
2015년 경기대학교 컴퓨터과학과(학사)
2015년~현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 컴퓨터비전

김 인 철

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