이미지 캡션 생성을 위한 심층 신경망 모델 학습과 전이

김동하 김인철
경기대학교 컴퓨터과학과
{kdh2040, kic}@kyonggi.ac.kr

Learning and Transferring Deep Neural Network Models for Image Caption Generation

Dong-Ha Kim Incheol Kim
Department of Computer Science, Kyonggi University

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 구성하고, 컨볼루션 신경망 층의 출력을 임베딩 층뿐만 아니라 멀티 모달 층에도 연결함으로써, 캡션 문장 생성을 위한 매 단계마다 이미지의 시각 정보를 이용할 수 있는 연결 구조를 가진다. Flickr8k, Flickr30k, MSCOCO 등의 공개 데이터 집합들을 이용한 다양한 비교 실험을 통해, 캡션의 정확도와 모델 전이의 효과 면에서 본 논문에서 제시한 멀티 모달 순환 신경망 모델의 우수성을 입증하였다.

1. 서론

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, (그림 1)에는 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 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 줄 것으로 기대하고 있다.

(그림 1) 이미지 캡션의 예들

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 구

* 본 연구는 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구과제 (No. 10060086, 개인 서비스용 로봇을 위한 지능-지식 집약, 개방, 진화형 로봇지능 소프트웨어 프레임워크 기술 개발)입니다.

성된다. 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 언어 모델이 처음 시작되는 임베딩 층(embedding layer)에만 연결하는 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), GRU(Gated Recurrent Unit) 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.

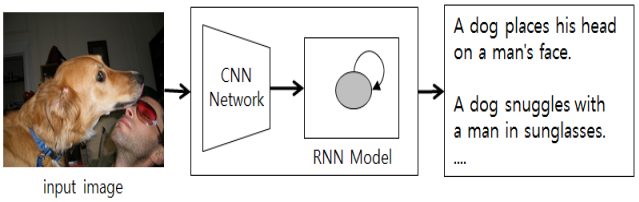
본 논문에서는 앞서 언급한 질문들에 답하기 위해, 효과적인 이미지 캡션 생성을 위한 멀티 모달 순환 신경망 모델을 제시한다. 본 모델에서는 시퀀스 패턴 학습과 모델 전이에 우수한 LSTM 유닛들로 순환 신경망 층(RNN layer)을 구성하며, 컨볼루션 신경망 층(CNN layer)을 통해 추출되는 시각 정보들을 매번 다음 단계 캡션 단어를 예측하는데 이용할 수 있도록 임베딩 층(embedding layer)뿐만 아니라 멀티 모달 층(multimodal layer)에도 연결하는 구조를 가진다. Flickr8k, Flickr30k, MSCOCO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

2. 관련 연구

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 정보들은 디코더 순환 신경망(decoder RNN)의 첫 단계 입력으로만 제공하는 연결 구조를 가지고 있다. 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 신경망 층(CNN layer)을 이용하며, 순환 신경망 층(RNN layer)은 LSTM 유닛들로 구성하였다. 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.

3. 이미지 캡션 생성

본 논문에서는 (그림 2)와 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.

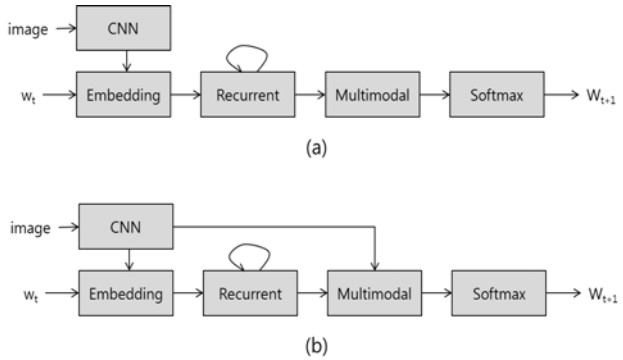


(그림 2) 이미지 캡션 생성을 위한 개념 모델

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 (그림 3)과 같이 서로 다른 시각 정보 연결 구조를 가질 수 있다.

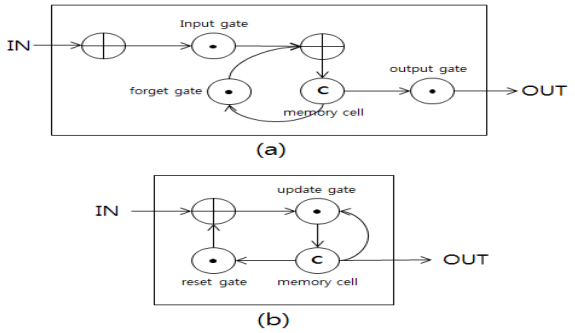


(그림 3) 시각 정보 연결 구조

(그림 3)의 (a)는 컨볼루션 신경망 층(CNN layer)의 시각 정보 출력을 임베딩 층(embedding layer)에만 연결하는 구조를 나타낸다. 이러한 시각 정보 연결 구조를 사용할 경우, 이미지의 시각 정보는 캡션 생성을 위한 순환 신경망의 첫 단계에 한번만 사용된다. 반면에, (그림 3)의 (b)는 컨볼루션 신경망 층(CNN layer)의 시각 정보 출력을 임베딩 층(embedding layer)뿐만 아니라 멀티 모달 층(multimodal layer)에도 연결하는 구조를 나타낸다. 이러한 시각 정보 연결 구조를 사용할 경우, 이미지의 시각 정보는 캡션 단어를 생성하는 순환 신경망의 매 단계마다 사용된다. 시각 정보를 캡션 단어를 생성하는 매 단계마다 사용하는 (그림 3)의 (b)의 연결 구조는 모델 학습과 캡션 생성에 소요되는 시간은 증가할 수 있으나, 캡션의 정확도는 훨씬 더 개선될 수 있을 것으로 판단한다. 따라서 본 논문에서는 (그림 3)의 (b)와 같은 연결 구조를 갖는 멀티 모달 순환 신경망 모델을 사용한다. 그리고 이미지로부터 시각 특징들을 추출하기 위해서 Inception v4 컨볼루션 신경망(CNN)을 이용한다.

3.2 순환 신경망 유닛

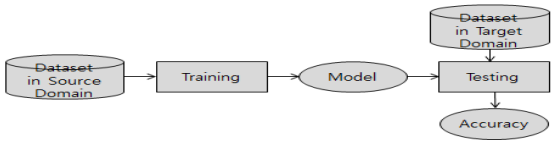
가중치 소멸 문제(vanishing gradient problem)을 극복하기 위해 새로 개발된 대표적인 순환 신경망 유닛들로는 LSTM과 GRU 등이 있다. (그림 4)의 (a)와 (b)는 각각 LSTM 유닛과 GRU 유닛의 내부 구조를 나타낸다. 하나의 LSTM 유닛은 (그림 4)의 (a)와 같이 입력 게이트(input gate), 출력 게이트(output gate), 망각 게이트(forget gate) 등 총 세 개의 게이트들로 셀 갱신(cell update)과 출력(output) 제어가 가능한 하나의 메모리 셀(memory cell)을 나타낸다. LSTM 유닛은 게이트들을 포함해 많은 수의 내부 파라미터들(parameters)을 포함하고 있어서, 많은 훈련 데이터와 긴 학습 시간을 요구하지만, 비교적 정확한 캡션 생성 모델을 얻을 수 있다. 반면에 하나의 GRU 유닛은 (그림 4)의 (b)와 같이 갱신 게이트(update gate), 리셋 게이트(reset gate) 등 단 두 개의 게이트로 셀 갱신과 출력을 조절할 수 있는 메모리 셀이다. 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 멀티 모달 순환 신경망 모델에서는 학습 시간 면에서 약간 유리한 GRU 유닛 대신 캡션 생성 모델의 정확도가 높은 LSTM유닛을 선택하였다.



(그림 4) LSTM과 GRU의 내부 구조

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.



(그림 5) 모델 전이

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 (그림 5)와 같이 원래 도메인(source domain)과 목표 도메인(target domain)의 다양한 변화에 대한 모델 전이 실험들을 수행한다.

4. 실험 및 평가

본 논문에서는 성능 실험을 위해 Flick8k, Flickr30k, MSCOCO 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. MSCCCO 의 경우는 훈련

데이터 82,783개, 테스트 데이터 40775개를 사용하였다. 각 이미지에는 다섯 문장 이상의 캡션이 함께 제공된다.

<표 2> Flickr8k의 캡션 정확도

	BLEU_1	BLEU_2	BLEU_3	BLEU_4
(a) - GRU	0.564	0.378	0.245	0.156
(b) - GRU	0.589	0.404	0.269	0.172
(a) - LSTM	0.582	0.395	0.261	0.168
(b) - LSTM	0.595	0.408	0.273	0.178

<표 2> Flickr30k의 캡션 정확도

	BLEU_1	BLEU_2	BLEU_3	BLEU_4
(a) - GRU	0.587	0.392	0.257	0.168
(b) - GRU	0.612	0.423	0.286	0.192
(a) - LSTM	0.604	0.410	0.274	0.184
(b) - LSTM	0.619	0.426	0.287	0.193

<표 3> MSCOCO의 캡션 정확도

	BLEU_1	BLEU_2	BLEU_3	BLEU_4
(a) - GRU	0.679	0.502	0.367	0.272
(b) - GRU	0.684	0.502	0.364	0.266
(a) - LSTM	0.674	0.493	0.356	0.259
(b) - LSTM	0.683	0.502	0.369	0.274

첫 번째 실험에서는 (그림 3)의 (a)와 (b)에 제시된 서로 다른 시각 정보 연결 구조와 LSTM, GRU 등 서로 다른 순환 신경망 유닛에 따른 캡션의 정확도와 모델 학습 시간을 비교해 보았다. <표 1>, <표 2>, <표 3>은 시각 정보 연결 구조 {(a), (b)}와 유닛 종류 {GRU, LSTM}의 서로 다른 네 가지 조합들에 따른 캡션 정확도(caption accuracy)를 평가한 실험 결과를 나타낸다. 캡션 정확도는 (식 1)과 (식 2)에 정의된 N 그램 문장 단위 평가 척도인 BLEU-N 계산식을 이용하여 평가하였다. (식 2)에서 r은 정답인 문장 수를, c는 생성된 문장 수를 나타낸다.

$$BP = \min(1, e^{1-\frac{r}{c}}) \qquad \text{(식 1)}$$

$$BLEU-N = BP \cdot e^{\frac{1}{N}\sum_{n=1}^{N}\log(p_n)} \qquad \text{(식 2)}$$

실험 결과, 본 논문에서 제안한 (b) 연결 구조와 LSTM 유닛의 조합((a)-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) 연결 구조에 비해 (b) 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.

<표 4> 모델 학습 시간

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

<표 4>는 시각 정보 연결 구조 {(a), (b)}와 유닛 종류 {GRU, LSTM}의 서로 다른 네 가지 조합들에 따른 모델 학습 시간(model learning time)들을 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. 순환 신경망 유닛만 비교해보면, 예상한대로 전반적으로 LSTM 유닛의 학습 시간이 GRU의 경우에 비해 좀 더 긴 것을 확인할 수 있다.

<표 5> GRU의 모델 전이 결과

Training \ Test	Flickr8k	Flickr30k	MSCOCO
Flickr8k		0.555	0.454
Flickr30k	0.619		0.496
MSCOCO	0.536	0.542	

<표 6> LSTM의 모델 전이 결과

Training \ Test	Flickr8k	Flickr30k	MSCOCO
Flickr8k		0.560	0.456
Flickr30k	0.620		0.493
MSCOCO	0.540	0.543	

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. <표 5>와 <표 6>는 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. 두 표에 제시된 모델 전이 결과는 BLEU_1로 측정한 캡션 정확도이다. 실험 결과에서, Flickr30k를 훈련 데이터 집합으로, MSCOCO를 테스트 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 확인해 주는 결과로 볼 수 있다.

5. 결론

본 논문에서는 이미지 캡션 생성에 효과적인 심층 신경망 모델을 제시하였다. 본 모델은 멀티 모달 순환 신경망 모델의 하나로서, 순환 신경망 층(RNN layer)은 시퀀스 패턴 학습과 모델 전이에 우수한 LSTM 유닛들로 구성되며, 시각 정보를 제공하는 컨볼루션 신경망 층(CNN layer)의 출력은 임베딩 층(embedding layer)과 멀티 모달 층(multimodal layer)에 모두 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.

참고문헌

[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, Alexander Toshev, et al., "Show and Tell: A Neural Image Caption Generator," Proc. of the IEEE, Conf. on CVPR. 2015.

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

[4] Junhua Mao, 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.