LSTM을 이용한 한국어 이미지 캡션 생성

박성재○, 차정원
창원대학교
tjdwo1289@gmail.com, jcha@changwon.ac.kr

Generate Korean image captions using LSTM

Seong-Jae Park○, Jeong-Won Cha
Changwon National University

요 약

본 논문에서는 한국어 이미지 캡션을 학습하기 위한 데이터를 작성하고 딥러닝을 통해 예측하는 모델을 제안한다. 한국어 데이터 생성을 위해 MS COCO 영어 캡션을 번역하여 한국어로 변환하고 수정하였다. 이미지 캡션 생성을 위한 모델은 CNN을 이용하여 이미지를 512차원의 자질로 인코딩한다. 인코딩된 자질을 LSTM의 입력으로 사용하여 캡션을 생성하였다. 생성된 한국어 MS COCO 데이터에 대해 어절 단위, 형태소 단위, 의미형태소 단위 실험을 진행하였고 그 중 가장 높은 성능을 보인 형태소 단위 모델을 영어 모델과 비교하여 영어 모델과 비슷한 성능을 얻음을 증명하였다.

주제어: 이미지 캡션 생성, Deep Learning, LSTM, CNN

1. 서론

스마트폰과 각종 센서들의 상용화로 인해 이미지 데이터의 양이 증가함에 따라 이미지 데이터의 활용성이 증가하고 있다. 그 중 이미지 캡션 생성기술은 이미지를 설명하는 텍스트를 생성하는 기술로 이미지에서 객체 인식 및 자연어 처리 기술에서 문장 생성의 기술을 필요로 하는 기술이지만 영어권에서는 이미 MicroSoft에서 시각장애인들을 위해 주변 환경을 설명해주는 씨잉 AI 앱[1]을 출시하는 등 상용화를 앞두고 있다.

그러나 한국어 이미지 캡션 생성기술의 경우 기계학습이나 딥러닝을 적용할 만한 한국어 이미지 캡션 데이터가 부족하다.

본 논문에서는 영어 이미지 캡션 데이터를 번역하여 한국어 데이터를 생성하였다. 딥러닝을 사용할 때 한국어 특성에 맞는 캡션 모델을 찾기 위해서 어절별, 형태소별, 의미형태소별 실험을 통해서 최적의 생성 모델을 찾았다.

2. 관련 연구

이미지 캡션 생성에 대해서는 다음과 같은 연구가 있었다. Multi-modal RNN을 이용한 방법[2]에서는 이미지 모델과 언어모델 그리고 두 모델을 통합하는 통합모델 총 3가지 모델을 이용해 이미지 캡션 생성을 진행하였다. 그리고 2015 Image Captioning Challenge에서 구글이 제안한 방법[3]은 CNN과 LSTM을 이용해 이미지 캡션을 생성하였는데 CNN은 Inception V3를 이용하였다. 또한 최근 CNN과 SVM을 함께 사용해 이미지 처리에 높은 성능을 보인 R-CNN과 RNN을 이용한 방법[4] 등 다양한 방법이 연구되었다. 한국어 이미지 캡션 생성은 CNN과 LSTM RNN의 변형인 GRU를 이용하는 방법[5]이 제안되었는데 이 연구에서는 Flickr 8K 데이터를 대상으로 영어 이미지 캡션을 번역자가 번역해 사용하였다. 따라서 [5]에서는 번역자가 직접 영어 데이터를 번역해야하는 단점이 존재한다.

3. 제안 방법

한국어 캡션 데이터를 생성하기 위해서 기존 영어 캡션 데이터를 번역을 이용해 한국어로 변환하였다.

영어권에서는 단어 단위를 입력으로 사용하여 캡션을 생성하였다. 하지만 한국어 모델의 경우 어절 단위로 입력을 사용하게 되는 경우 조사에 따라 같은 단어도 다른 단어로 인식하는 문제가 존재한다. 따라서 기존 데이터에 비해 학습데이터의 양이 훨씬 많아야 하는 문제가 있다.

어절 단위 학습데이터와 형태소분석을 거친 형태소 단위 학습데이터를 생성하였고 추가적으로 문장의 의미를 생성하는 것은 기능형태소가 아닌 의미형태소라고 판단되어 의미형태소만 남긴 데이터를 구축하였다.

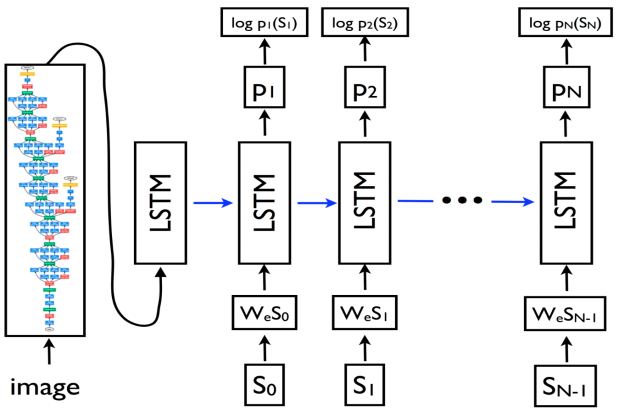

그림 1 이미지 캡션 생성 모델 구조도

본 논문에서는 그림 1과 같이 CNN과 LSTM을 사용한 이

미지 캡션 생성 모델[3]을 사용한다. 입력 이미지를 CNN을 이용하여 512차원의 자질로 인코딩하며 인코딩된 자질을 이용하여 문장을 생성하는 LSTM으로 구성되어있다. CNN은 이미지 구글의 Inception V3를 사용하여 마지막 히든 레이어의 값을 LSTM의 입력으로 사용하였다.

4. 실험 및 토의

본 논문에서는 한국어 이미지 캡션 생성을 위해 MS COCO데이터 셋을 사용하였다. MS COCO 데이터 셋은 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.

제안 방법과 같이 생성된 어절 단위, 형태소 단위, 의미형태소 단위 학습데이터를 이용해 step을 각 200,000번으로 설정하여 학습 후 실험을 진행하였다.

성능평가에는 기계번역에서 성능지표로 사용하는 BLEU score를 사용하였다.

표 1은 학습 데이터 유형별 실험 결과이다. 각 모델별 실험 결과를 비교했을 때 형태소 단위의 실험이 가장 성능이 높고 어절 단위 실험이 가장 성능이 낮은 것을 확인할 수 있다.

표 1 학습데이터 유형별 실험결과

Model	B-1	B-2	B-3	B-4
어절 단위	0.289	0.190	0.141	0.111
의미형태소 단위	0.597	0.392	0.286	0.225
형태소 단위	0.615	0.430	0.322	0.251

46. 이창수, 천주룡, 김주근, 김태일, 강인호, Naver Search, "Distance LSTM-CNN with Layer Normalization을 이용한 음차 표기 대역 쌍 판별"

가장 낮은 성능을 보인 어절 단위 모델의 경우 조사에 따라 동일 단어가 다른 단어로 인식하기 때문에 데이터의 부족으로 인한 오류가 많이 발생한 것으로 생각된다. 형태소 단위 모델이 의미형태소 단위 모델보다 성능이 좋은 것은 기능형태소가 제한적인 어휘를 가지기 때문에 예측 성능이 높게 나타나기 때문이라고 생각된다. 이를 증명하기 위한 추가 실험으로 형태소 단위 실험결과의 정답과 예측 데이터를 각각 의미형태소와 기능형태소만 남기고 성능을 측정하였다. 표 2의 결과를 보면 기능 형태소의 성능이 높은 것을 확인할 수 있다.

표 2 형태소 단위 모델의 세부 실험 결과

Model	B-1	B-2	B-3	B-4
의미형태소	0.547	0.366	0.280	0.230
기능형태소	0.686	0.457	0.346	0.279

따라서 이 중 가장 높은 성능을 보인 형태소 단위 모델을 step을 400,000번으로 증가시켜 학습을 진행하고 영어 모델은 step을 1,000,000번으로 증가 시켜 학습해 두 모델의 성능을 비교하였다.

표 3은 형태소단위, 영어 모델의 실험결과와 MS COCO Image Captioning Challenge에서 높은 성능을 보인 5개의 모델의 결과이다. 400,000번 학습하여 생성한 형태소 단위 모델의 성능이 학습량이 적음에도 불구하고 영어 모델의 성능보다 B-4 score로 0.100 더 높음을 확인하였으며 나머지 MS COCO Image Captioning Challenge에서 높은 성능을 보인 모델과 비교하더라도 크게 낮지 않은 성능을 보임을 확인하였다.

표 3 실험 결과

Model	B-1	B-2	B-3	B-4
형태소 단위	0.630	0.445	0.333	0.260
영어	0.611	0.441	0.323	0.250
NIC[3] (google)				0.309
MSR Captivator[6]				0.308
m-RNN(2) [2]				0.302
m-RNN [2]				0.299
MSR [7]				0.291

그림 2와 3은 출력결과의 예를 보여준다.

(가) (나) (다)

그림 3 MS COCO 평가 이미지 중 올바른 결과의 예

그림 2를 대상으로 형태소 모델과 영어 모델이 생성한 캡션은 가)의 경우 “야구선수가 공을 치려고 합니다.” 와 “a group of people on a field paying baseball.” 로 형태소 모델과 영어 모델 모두 이미지를 잘 설명하는 캡션을 생성한 것을 확인할 수 있다. 나)의 경우 “고양이는 나무벤치에 앉아있다.” 와 “a cat sitting on top of a wooden bench.” 를 캡션으로 생성하였고 다)의 경우 “바나나의 큰 무리가 시장에 있습니다.” 와 “a market with a bunch of bananas hanging from the ceilling” 을 생성하였다. 그림 2의 세 개의 이미지를 대상으로 형태소 모델과 영어 모델이 생성한 캡션이 이미지를 잘 설명하고 있는 것을 확인 할 수 있다.

(가) (나) (다)

그림 4 MS COCO 평가 이미지 중 오류 결과의 예

그림 3을 대상으로 형태소 모델과 영어 모델이 생성한 캡션은 가)의 경우 “케이크를 절단하는 칼을 들고 여자.” 와 “a man and woman cutting a cake with a knife” 를 캡션으로 생성하였다. 그러나 가)의 정답은 “요리사는 초밥을 만드는 레스토랑에 있습니다” 로 형태소 모델의 경우 “요리사” 를 “여자” 로 , “초밥” 을 “케이크” 로 기술하였다. 이는 “요리사” 와 “초밥” 이 학습 코퍼스에 나타난 빈도가 적어 생기는 문제로 예측되며 실제로 생성된 단어 사전을 보면 “여자” 의 경우 37,482번 , “요리사” 의 경우 671번 발생하였다. 마찬가지로 “케이크” 의 경우 8,894번 발생하였으나 “초밥” 의 경우 62번 발생하였다.

나)의 경우 “테이블에 앉아있는 샌드위치.” 와 “a sandwich sitting on top of white plate” 를 생성하였는데 위와 동일하게 117번 나타난 “콜라” 를 인식하지 못하는 문제가 발생하였다. 다)의 경우 “프리즈비를 들고 잔디에 서있는 어린 소년” 과 “a man in a field with a frisbee” 를 생성하는 등 학습 코퍼스에 저빈도로 나타난 사물을 오 인식하여 잘못된 캡션을 생성하는 문제가 발생하였다.

이렇듯 캡션이 부정확하게 생성되는 경우는 학습데이터의 부족으로 인한 오인식이며 이를 해결하기 위해서는 추가적인 학습데이터를 사용해야 할 것이라고 생각된다.

5. 결론 및 향후연구

본 논문에서는 한국어 이미지 캡션을 생성하기 위해 영어 캡션데이터를 번역을 이용해 한국어로 변환하는 방법을 제안하였다. 또한 어절단위보다 형태소단위로 학습을 진행하는 경우 성능이 더 높음을 보였다.

향후연구로는 LSTM의 입력으로 사용되는 Inception V3가 아닌 VGGNet[8] 또는 이미지 처리에 높은 성능을 보이고 있는 Fast R-CNN등을 사용하는 모델을 적용할 예정이다. 또한 의미형태소로 생성된 캡션을 seq2seq를 이용하여 실제 문장을 생성하는 연구를 진행할 계획이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음.[2015-0-00219, 개방형 미디어 생태계 구축을 위한 시맨틱 클러스터 기반 시청상황 적응형 스마트방송 기술 개발]

참고문헌

[1] https://www.microsoft.com/en-us/seeing-ai/
[2] MAO, Junhua, et al. Deep captioning with multimodal recurrent neural networks (m-rnn). arXivpreprint 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, 2017, 39.4: 652-663
[4] KARPATHY, Andrej; FEI-FEI, Li. Deep visual-semantic aligments for generating image descriptions. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 3128-3137
[5] 배장성, 이창기. (2016). 딥러닝을 이용한 한국어 이미지 캡션 생성. 한국정보과학회 학술발표 논문집. , 488-490
[6] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. “Language models for image captioning: The quirks and what works,” in ACL, 2015.
[7] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J.Platt, C. L. Zitnick, and G. Zweig, “From captions to visual concepts and back,” in CVPR, 2015.
[8] SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.