|  |
| --- |
| **Highway LSTM 디코더를 이용한**  **이미지 캡션 생성** |
| 박찬민  서강대학교  cmPark@sogang.ac.kr |
| Image Caption with Highway LSTM Decoder |
| Park Chan Min  Sogang University |
| **요 약**  딥 러닝 기술은 이미지 인식, 자연어처리 및 음성인식 등에서 다양하게 활용되고 있다. 이미지 캡션은 컴퓨터가 이미지를 이해하고 자연어로 변환하는 작업으로써 이미지 인식과 자연어처리 기술이 필요한 작업이다. 이미지 캡션 모델은 이미지를 벡터화하는 CNN 기반 인코더와 자연어를 생성하는 RNN 기반 디코더로 구성된다. 본 논문에서는 딥 러닝의 문제점 중 하나인 그래디언트 소멸 문제 (vanishing gradient problem)를 해결하기 위해 디코더에 Highway LSTM을 적용하였다. 실험 결과로 본 논문에서 제안하는 모델이 기존 모델보다 BLEU 점수가 0.4% 향상 되었음을 확인 할 수 있다. | | |

**1. 서 론**

이미지 캡션은 컴퓨터가 이미지를 이해하고 이해한 내용을 바탕으로 자연어로 문장을 생성하는 작업을 말한다. 이를 처리하기 위해서는 이미지 인식과 자연어처리에 관한 기술과 지식이 필요하다. 최근 딥 러닝을 이용한 이미지 인식과 자연어처리 연구가 활발히 진행되면서 자연스럽게 이미지 캡션 분야도 많은 연구가 진행되었다. 컴퓨터는 이미지를 이해하기 위해 CNN을 이용하여 이미지를 고차원 벡터로 변환한다. 변환된 벡터는 자연어로 변환하기 위해 RNN 기반의 디코더의 입력으로 사용된다.

이미지 인식과 자연어처리에서 사용되는 딥 러닝은 모델은 사람이 직접 자질(feature)을 추출 하지 않아도 기계학습 모델에 비해 성능이 높다는 장점이 있다. 하지만 대용량 학습이 데이터가 필요하다는 점과 모델의 레이어가 깊어 질수록 그래디언트 소멸 문제가 발생해 학습이 잘 되지 않는다는 단점이 있다. 본 논문에서는 그래디언트 소멸 문제를 부분적으로 해결하기 위해 Highway LSTM로 디코더를 구성했다.

본 논문에서 제안하는 모델의 CNN 인코더는 [1]에서 제안하는 모델과 동일하게 VGGnet[2]을 사용했다. 마찬가지로 디코더에서 자연어를 생성 할 때 이미지의 문맥정보(context)을 활용하는 soft-attention을 사용하였다.

실험결과 제안하는 모델은 기존 LSTM 디코더를 사용한 모델보다 0.4% 높은 성능을 보였다.

**2. 관련 연구**

최근 딥 러닝 기반의 이미지 인식 연구와 인코더-디코더 방식의 기계 번역 모델[3]의 영향을 받아 다양한 이미지 캡션 경진대회가 개최되었다. [4]에서는 VGGnet 기반 인코더 모델과 RNN기반 디코더 모델을 사용하여 우수한 성적을 냈다. [1]에서는 기계 번역 등에서 사용되는 attention 메커니즘을 이미지에 적용하여 향상된 성능을 보였다.

그래디언트 소멸 문제는 딥 러닝의 단점 중 하나이다. 네트워크의 구조가 복잡해짐에 따라 그래디언트 소멸 문제는 점차 중요해지고 있다. 이미지 인식에서 이를 해결하기 위해 [5]에서는 Residual Network를 제안했다. 비선형 변환을 거친 정보와 비선형 변환을 거치지 않은 정보를 학습을 통한 일정 비율로 사용한다. 비선형 변환을 거치지 않은 정보가 그래디언트 소멸 문제를 부분적으로 해결한다는 장점이 있다. 자연어처리 연구에서도 그래디언트 소멸 문제를 해결하기 위해 Residual Network와 유사한 모델을 사용하였다. [6]에서는 LSTM과 Highway Network를 결합한 모델을 사용하여 의미역 결정에서 우수한 성능을 보였다.

본 논문에서는 [1]에서 제안하는 soft-attention 모델과

[6]의 Highway LSTM을 결합한 모델을 제안한다.

**3. 제안 모델**

그림 1은 본 논문에서 제안하는 모델의 시스템 구조도이다. 먼저 입력 이미지는 CNN 기반의 VGGnet을 통해 고차원 벡터로 변환된다. 변환된 이미지 벡터는 Batch Normalization을 거친 뒤 Highway LSTM으로 구성된 디코더의 입력으로 사용된다. 디코더는 매 단어를 생성할 때 이미지로부터 일종의 문맥정보를 받기 위해 soft-attention 기법을 사용한다.

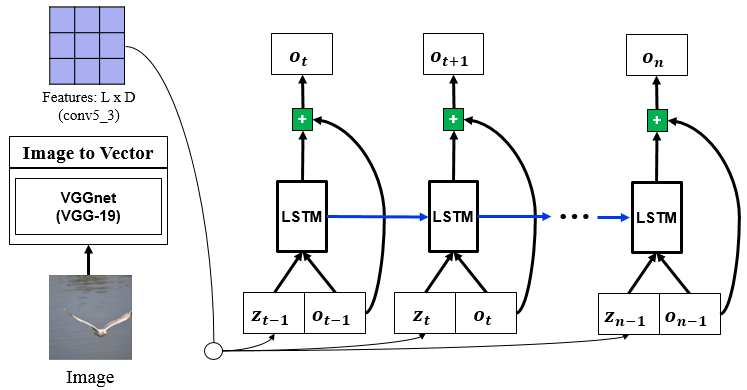


그림 1 제안 모델의 시스템 구조

3.1 CNN 인코더

본 논문에서는 ImageNet Large Scale Visual Recognition Challenge(ILSVRC) 2014[7]에서 우수한 성능을 보인 VGG Net을 인코더로 사용했다. 그림 2는 VGGnet의 구조이다.

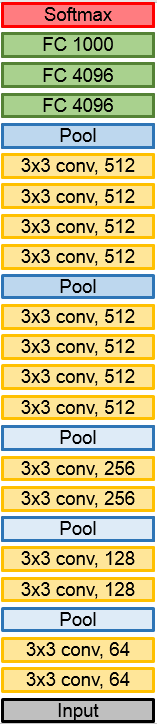


그림 2 VGGnet (VGG19) 구조

VGGnet은 이전 ILSVRC 2012에 제안되었던 AlexNet[8]에 비해 작은 크기인 3x3 CNN 필터를 사용한다는 특징이 있다. 2개의 연속된 3x3 CNN 필터는 1개로 구성된 7x7 필터와 동일한 수용면적을 갖는다. 이를 통해 더 깊은 계층의 모델을 구성하여 향상된 성능을 얻을 수 있다. 또한 더 적은 파라미터로 모델을 구성 할 수 있다는 장점이 있다.

입력 이미지는 VGGnet 기반의 인코더를 통해 고차원 벡터로 변환된 후 Batch Normalization 계층을 거쳐 디코더로 전달된다.

3.2 Highway LSTM(Long Short Term Memory)

LSTM은 그래디언트 소멸문제를 해결한 딥 러닝 모델로써 식 (1)과 같이 정의된다.

(1)

는 LSTM의 memory cell gate 벡터이며 각 벡터의 크기는 LSTM 의 은닉 계층인 와 동일한 크기를 갖는다.

Highway Network는 각 계층의 활성화 함수를 거쳐 비선형 변환이 된 입력 정보와 비선형 변환을 거치지 않은 정보를 더한 결과를 다음 계층에서 사용하는 모델이다. 본 논문에서는 입력 계층과 LSTM의 은닉 계층 사이에 Highway Network를 적용하였다. LSTM의 입력은 3.3에서 설명할 attention 와 이전 생성 단어 가 합쳐진 입력 벡터이다. 식 (2)는 Highway LSTM 기반의 디코더 수식이다.

(2)

는 transform gate로써 비선형 변환을 거친 정보와 비선형 변환을 거치지 않은 정보를 얼마나 사용할지 학습에 의해 결정하는 역할을 한다. 이러한 구조로 인해 역전파 단계에서 그래디언트가 소멸되지 않고 전파 된다는 장점이 있다. 는 출력 계층을 거쳐 현재 단어를 출력한다.

3.3 Soft Attention

본 논문에서는 [1]에서 사용된 Soft-Attention을 사용했다. 입력 이미지는 시각적인 정보를 의미하는 L개의 자질 벡터. 로 변환되어 디코더에 입력으로 사용된다. 각 자질 벡터 는 이미지의 부분적인 시각적 정보를 의미한다. 본 논문에서는 인코더 VGGnet의 5번째 CNN 계층의 마지막 필터(conv5\_3)를 사용하였다. Soft-Attention은 기계 번역에서 사용되는 Attention 메커니즘[9]과 유사하다. 식 (3)은 본 논문에서 사용되는 Soft-Attention의 수식이다.

(3)

는 attention은 학습 파라미터에 의한 별도의 Attention 계층이다. 는 현재 단어 를 생성하는데 이미지의 어느 부분에 집중할 것인가에 대한 가중치를 의미한다. 은 이전 생성단어 과 결합되어 디코더의 입력 벡터로 사용된다.

**4. 실험 및 결과**

표 (1)은 본 논문에서 제안하는 이미지 캡션 모델의 실험 결과이다. 비교 모델은 [1]에서 제안하는 soft-attention 모델이다. 본 논문에서는 총 3번에 걸쳐 기존 모델과 제안 모델을 비교실험 하였다. 제안 모델의 성능이 기존 모델에 비해 0.4% 향상된 것을 알 수 있다.

|  |  |  |
| --- | --- | --- |
|  | **기존 모델** | **Highway LSTM** |
| BLEU-1 (1) | 65.99% | 66.28% |
| BLEU-1 (2) | 66.01% | 66.49% |
| BLEU-1 (3) | 66.16% | 66.60% |
| 평균 | 66.06% | **66.46%** |

표 1 실험 결과

**5. 결 론**

본 논문에서는 [1]에서 제안한 이미지 캡션 모델과 [6]의 Highway LSTM 모델을 결합한 Highway LSTM 디코더 기반의 이미지 캡션 모델을 제안한다. Highway LSTM은 그래디언트 소멸 문제를 부분적으로 해결한 모델로 이미 의미역 결정, 개체명 인식과 같은 자연어처리 연구에서 향상된 성능을 보였다. 마찬가지로 이미지 캡션에서도 성능 향상이 있음을 본 논문의 실험을 통해 알 수 있다. 향후 계획으로 인코더 모델을 파인 튜닝하는 과정과 사전 학습된 단어 임베딩 등을 추가할 예정이다.

**참고문헌**

[1]Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International Conference on Machine Learning*. 2015.

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[3] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

[4] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[6] He, Luheng, et al. "Deep semantic role labeling: What works and what’s next." *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2017.

[7]<http://www.image-net.org/challenges/LSVRC/2014/>

[8] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[9] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, September 2014.