Encoder- Decoder 아키텍쳐 종류에 따른 이미지 캡션 분석

배나연, 한동석*
경북대학교 대학원 전자전기공학부

qoskdus1@kau.ac.kr, *dshan@knu.ac.kr

Image caption performance analysis according to Encoder-Decoder Architecture type

Bae Na Yeon, Dong Seog Han*
School of Electronic and Electrical Engineering, Kyungpook National Univ.

요 약

이미지 캡션 생성은 이미지를 설명하는 문장을 자동으로 생성하는 기술을 의미한다. 이미지 캡션 모델은 이미지의 특징을 추출하는 Encoder 와 추출된 특징을 자연어로 처리하여 주는 Decoder 로 구성되어 있다. 본 논문은 Encoder 에서 쓰는 CNN 아키텍처와 Decoder 에서 사용하는 RNN 아키텍처를 결합하여 종류별로 캡션 생성을 하고 생성된 캡션을 평가하고 각 아키텍처 별로 이미지 캡션 분석을 한다.
Keywork : image caption, Convolution Neural Network, Recurrent Neural Network

Ⅰ. 서 론

이미지 캡션 생성은 이미지를 설명하는 문장을 자동으로 생성하는 기술을 말한다[1]. 이미지 캡션 모델은 컴퓨터 비전과 관련된 작업에서 물체를 인식하고, 물체들 간의 관계를 파악하여 자연어로 표현할 수 있는 자연어 처리 능력을 가질 수 있어야 한다. 이미지 캡션은 이미지의 특징을 추출하여서 이미지의 특징을 자연어 문장으로 번역하는 번역의 일종으로 볼 수 있다.

이미지 캡션 생성의 모델은 Encoder 와 Decoder 로 구성되어 있다. Encoder 에서는 Convolution Neural Network(CNN)를 사용하여 이미지가 가진 특징을 추출하고 Decoder 에서 Encoder 에서 추출된 이미지의 특징을 Recurrent Neural Network(RNN)을 사용하여 캡션을 생성할 수 있다.

본 논문에서는 ResNet101[2], VGG-16[3 ,GoogleNet[4]과 같은 Convolution Neural Network(CNN)와 GRU, LSTM 과 같은 Recurrent Neural Network(RNN)를 결합하여 이미지 캡션 생성하고 생성된 캡션을 평가하고 분석을 진행한다.

Ⅱ. 본론

본 논문은 ResNet101, VGG-16, GoogleNet 과 LSTM, GRU 아키텍처를 융합하여 만든 모델을 사용하여 만들어진 입력된 이미지에 이미지 캡션 생성하고 분석한다.

Neural Image Caption[1] 모델을 사용하여 그림 1 과 같이 모델을 만들었다.

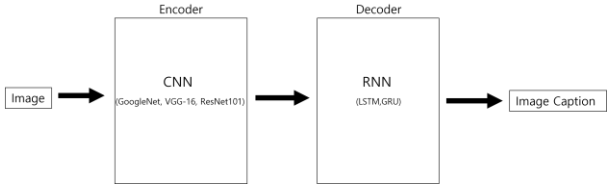

그림 1. Image Caption Model

모델의 아키텍처를 변화시켜서 총 6 개의 모델을 만들고 학습시켰다. 학습의 위한 데이터셋으로 flickr8k 데이터셋을 사용하였고, batch 사이즈 64, epoch 10 으로 학습을 시켰다.

그림 2. Flickr8k 개가 나무를 뛰어 넘는 사진

그림 3. Flickr8k 사람들이 골짜기를 구경하는 사진

VGG-16 & LSTM	Dog is running through the grass Man stands on rock overlooking the mountains
VGG-16 & GRU	Dogs play in the grass Man in red shirt is standing on rock overlooking valley
GoogleNet & LSTM	Dog is running through the woods man in red shirt is standing on rock overlooking the mountains
GoogleNet & GRU	Dog is jumping over fallen tree People are sitting on the edge of mountain
ResNet101 & LSTM	Dog is jumping over tree Main in red shirt is standing top of mountain
ResNet101& GRU	Black dog jumps over log Man is standing on top of mountain overlooking valley

표 1. 이미지 캡션 생성 결과

그림 2 와 3 에 이미지 캡션을 생성하였을 때의 결과는 표 1 과 같다. 그림 2 는 객체 하나에 초점을 둔 그림인데,

이러한 그림에서는 모든 모델이 비슷한 이미지 캡션 생성을 보여준다. 그림 3 은 여러 객체가 나오는 사진인데, 대부분의 모델이 객체 하나에 초점을 두어 이미지 캡션 생성을 하였다. 하지만 GoogleNet 과 GRU 아키텍처를 쓴 모델의 경우에는 여러 객체 인식하여서 객체 간의 관계를 파악하는 결과를 보여주어 다른 아키텍처의 결합에 비해서 이미지의 특징의 관계를 더 잘 파악하여 이미지 캡션 생성하여 다른 모델에 비해서 우수한 성능을 보여주었다.

Ⅲ. 결론

본논문에서는 Encoder Architecture 와 Decoder Architecture 를 종류 별로 이미지 캡션 생성하여 각 모델들을 분석 비교하였다. 대부분의 모델들이 비슷한 이미지 캡션을 생성을 하는 것 보여주었고, GoogleNet-GRU 결합 모델은 다른 모델에 비해서 자세한 캡션 생성을 해주는 것을 보여주었다.

ACKNOWLEDGMENT

이 논문은 2023 년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임

(2021R1A6A1A03043144)

참 고 문 헌

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

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

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

[4] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.