행동 인식 참조 이미지 캡셔닝

박은수, 김승환, 유재성, 김선대, 굴람 무즈타바 류은석
가천대학교 컴퓨터공학과
{dmsehf804, whitekomani, poopoo96, ele7004, Mujtaba}@gc.gachon.ac.kr,
esryu@gachon.ac.kr

Action Recognition Reference Image Captioning

Eun-Soo Park, Seunghwan Kim, Jaesung Ryu, Seondae Kim, Ghulam Mujtaba, Eun-Seok Ryu
Computer Engineering, Gachon University

요 약

본 논문에서 기존의 이미지 캡셔닝의 문제점인 행동 인식 관련한 문제를 해결한다. 이미지 캡셔닝 모델의 학습 데이터의 행동 부분 즉, 동사 부분으로 행동 인식 데이터 셋을 만들었을 경우 많은 클래스, 각 클래스에는 적은 데이터로 구성됨을 보였다. 따라서, 본 논문에서 행동 인식 모델을 추가하고, 임계값을 두어 이미지 캡셔닝의 동사 부분의 정확도가 낮을 경우, 그리고 행동 인식 모델의 정확도가 높을 경우 두 결과물을 교체하는 방식으로 이미지 캡셔닝의 문제점을 해결한다. 본 논문에서 제안하는 모델에 대한 설명과 구현 과정 및 행동 인식에 강인한 이미지 캡셔닝 실험 결과를 보인다.

1. 서론

최근 고성능 GPU 의 사용으로 처리가능한 연산량이 대폭 증가함에 따라, 패턴을 인식하는데 연산량이 많이 필요한 딥 러닝 기술이 계속 연구 되고있다. 합성곱신경망(Convolution neural network, CNN)[1]과 같은 신경망의 발달과 함께 객체 인식, 이미지 분류 등과 같은 이미지 프로세싱 연구가 상당히 빠른속도로 진행되어 오고 있다. 헬스케어분야의 딥 러닝기술 적용으로인하여, 사람이 포함된 영상 이해, 상황 인식과 같은 연구가 심도있게 진행되어 오면서, 딥 러닝 기반의 영상 캡셔닝의 중요도가 부각되어 오고있다.

이미지 캡셔닝 (Image captioning)이란 입력된 이미지를 합성곱신경망을 통하여 특징을 추출하고, 학습된 단어 특징 공간에 매핑하여 입력된 이미지의 설명문을 생산하는 것으로, 영상 이해 및 상황 인식에 가장 근접한 연구 중 하나이다.

그림 1 과 같이 기존 이미지 캡셔닝 모델을 사용할 경우 한 장의 이미지로 행동을 예측하는 부분에서 큰 어려움이 있을 수 있다. 같은 행동이지만, 다르게 예측될 수 있다. 본 논문에서 Flickr 8k[2] 데이터 셋의 이미지의 설명 데이터에서 자연어 처리 모듈을 사용하여, 동사 부분만을 출력 후 데이터 셋을 만들었다. Flickr 8k 의 동사 부분만을 모아 만든 데이터 셋의 클래스의 개수는 총 1523 가지이다. 각 클래스별 이미지의 개수는 전반적으로 10 장 내외다. 이때 인칭 동사 (e.g. is, are, etc.)는 제외하였다. 이 실험으로 보아, 이미지 캡셔닝 데이터 셋으로 학습할 경우 행동 인식 관련 부분은 상당히 많은 클래스와 각 클래스의 이미지 개수는 적은 상태로 학습을 진행하는 것으로 볼 수 있다. 따라서, 본 논문에서 기존의 이미지 캡셔닝 모델로 정확하게 판단하기 힘든 행동 인식 부분은 선행학습된 행동 인식 모델을 사용하여, 이미지 캡셔닝의 정확도를 높인다.

본 논문은 앞서 언급한 이미지 캡셔닝의 문제점을 해결하기 위한 방법을 서술한다. 이에 대한 구성으로 2 절에서 본 논문의 관련연구로 이미지 캡셔닝, 행동 인식에 대해 기술되어 있고, 3 절에서 행동 인식 참조 이미지 캡셔닝 모델을 제안, 구현 및 데이터 셋에 대한 설명을 기술한다. 4 절에서 실험 결과 및 분석을 기술하고, 마지막으로 5 절에서 결론을 기술한다.

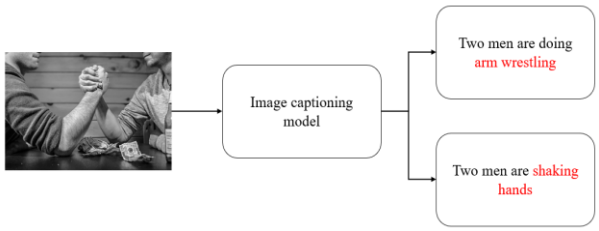

그림 1. 기존 이미지 캡셔닝 모델의 행동 인식 관련 문제점

2. 관련 연구

2 절에서 본 논문에서 이용된 기술인, 이미지 캡셔닝과 행동 인식 모델에 대한 지금까지의 연구들을 기술한다.

2.1 이미지 캡셔닝

이미지 캡셔닝은 언어 번역기의 아이디어에서부터 시작되었다. 인코딩과 디코딩 구조로 되어있는 번역기는 도메인(언어)에서 도메인 (언어)으로 변환을 하는 방식이다. 이미지 캡셔닝도 마찬가지로 도메인 (이미지)에서 도메인 (언어)로

변환을 하는 것이다. 이미지를 언어로 변환하기 위하여 이미지의 특징을 추출하여, 특징 벡터로 변환을 해야 한다. 이때, 영상의 특징을 추출하는데 특화된 합성곱신경망을 사용한다. 합성곱신경망을 통해 추출된 특징은 학습된 단어 임베딩 특징 벡터 공간에 연결한다 (Concatenate). 생성된 특징 공간을 GRU[3], LSTM[4]과 같은 순환 신경망(Recurrent Neural Network, RNN)에 입력한다. 순환 신경망은 순서 및 시간이라는 측면을 고려할 수 있기 때문에 이전에 생성된 단어를 참고하여 다음 단어를 생성하므로 정확도를 더 높일 수 있다[5].

초기의 이미지 캡셔닝의 성능을 높이기 위하여 여러가지 연구가 진행되었다. 디코더 부분에서, 현재 캡션을 출력할 때, 어느 영역의 가중치가 높은지, 또한 초점을 두고 있는 부분을 제외한 나머지의 가중치를 억제하는 집중 (Attend) 모델[6], 여러 객체가 있을 경우, 객체 인식 프로그램으로 여러 객체의 영역을 나누어, 각각의 객체에 대해 캡션을 진행하여 고밀도 동시 발생 객체 오류를 줄이는 연구가 있다[7].

2.2 행동 인식

행동 인식은 영상 내에서 인간의 행동을 분석하는 기술로 딥 러닝을 사용하였을 때 높은 효과를 보였다. 초기에는 다중 흐름 (Two-stream) CNN 기법으로 컬러 이미지와 행동의 특징을 나타낼 수 있는 데이터를 이용하여 행동인식을 하였다[8]. 행동 인식에 대한 연구는 데이터 전처리에 관한 연구가 더 많이 이루어지고 있다. 키넥트 센서를 이용하여 3d 포즈를 추정하고, 스켈레톤 이미지와 함께 입력하여 행동인식을 하는 연구 등이 있다[9].

3. 행동 인식 참조 이미지 캡션 모델

3 절에서 제안하는 행동 인식 참조 이미지 캡션 모델에 대한 설명과 모델 구현 방법 및 이용한 데이터 셋에 대한 것을 기술한다.

3.1 모델 개요

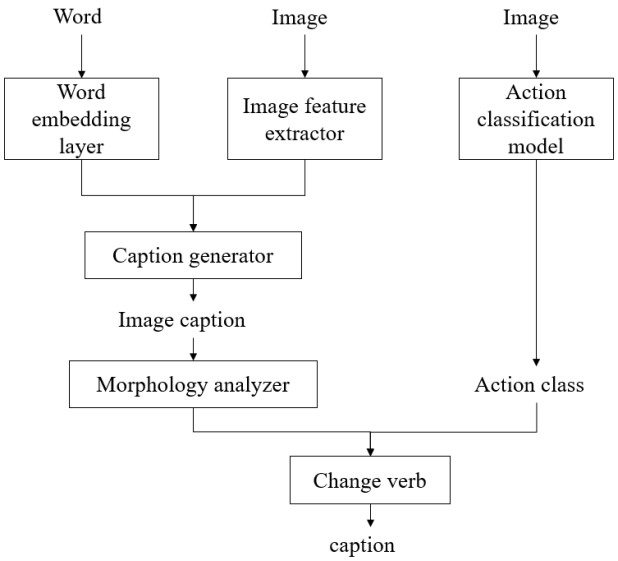

그림 2. 제안하는 행동 인식 참조 이미지 캡셔닝 모델

본 논문에서 제안하는 행동 인식 참조 이미지 캡셔닝 모델의 전체 구조는 그림 2 와 같다. 먼저 이미지 캡션 모델 부분은 합성곱신경망을 사용하여 이미지의 특징을 인코딩한다. 인코딩된 데이터는 단어 특징이 임베딩 되어있는 디코딩 단계에 입력된다. 디코딩 단계에서 입력된 특징 벡터 공간을 사용하여 단어를 출력한다. 이 때, 한 장의 이미지를 입력하여 행동, 상황 등 여러가지를 하나의 합성곱신경망으로 추정해야하므로, 정확도가 상당히 떨어질 수 있다. 따라서 행동 인식 데이터 셋으로 학습된 행동 인식 모델을 추가하여 부족한 부분을 보완한다. 부족한 부분은 동사 부분이므로, 동사 단어의 위치를 알기 위하여 이미지 캡셔닝에서 출력된 캡션을 품사 (part-of-speech, POS)분석한다. 제안하는 모델에서 생성되는 캡션의 손실 값으로 교체 여부를 판단한다. 이때, 동사의 손실 값이 아닌, 다른 품사의 손실 값이 낮아서 생기는 오류를 방지하기 위하여, 알고리즘 1 과 같은 과정을 거친다.

알고리즘 1

캡션의 동사 부분에 행동 인식 결과 값을 적용하는 과정

M_p: image captioning model prediction
$M_C(pos\_tag)$: part of speech of predicted caption
M_L: loss value of image captioning model
V: max length of caption
$V_{condition}$: verb condition (e.g. VB, VBD, etc.)
A_c: predicted action class
A_p: accuracy of action classification model
C: caption
for $i < V$ **do**
  M_C, $M_L \leftarrow M_p$ ($image$, cap)
  if $M_L <$(score/i) **then**
    if $M_C(pos\_tag)$ == $V_{condition}$ **then**
      if $A_p > 0.8$ **then**
        $C \leftarrow A_c$
      end if
    end if
  else
    $C \leftarrow M_C$
  end if
  i ← i + 1
end for

이미지 캡션의 품사 분석은 자연어 처리 모듈인 Natural language toolkit (NLTK)를 사용하였다. 품사 분석은 NLTK 의 pos_tag 메소드를 사용하여, 사전에 NLTK 에 입력된 단어와 매핑되어 있는 태그로 캡션내의 각 단어들에 태그를 넣을 수 있다. NLTK 내에 있는 동사관련 태그는 총 6 개로 표 1 과 같다. 행동 인식 모델의 클래스에 인칭 동사가 없으므로, VBP, VBZ 태그를 제외한 나머지 동사 태그들을 사용한다.

표 1. NLTK 동사 태그 종류

Tag	Description	Example
VB	base form	‘take’
VBD	past tense	‘took’
VBG	gerund/present participle	‘taking’
VBN	past participle	‘taken’
VBP	non-3rd person singular present	‘take’
VBZ	3rd person singular present	‘takes’

그림 3. 행동 인식 참조 이미지 캡셔닝 결과물

3.2 모델 구현

제안하는 모델의 학습은 종단간 학습 (end-to-end)으로 진행되지 않고, 이미지 캡셔닝 모델과 행동인식 모델을 각각 학습시키는 방식으로 진행되었다. 이미지 캡셔닝 모델의 경우 이미지의 특징 벡터를 추출하는 부분인 합성곱신경망은 이미지 데이터 셋 중 하나인 ImageNet 으로 선행 학습된 Inception V3[10] 모델을 사용하였다. 최적화 알고리즘은 Adam 을 사용하였고, 손실 함수는 수식 (1)과 같이 범주별 크로스엔트로피 (categorical cross-entropy)를 사용하였다.

$$L_{categorical} = -\frac{1}{n}\sum_x [y log\tilde{y} + (1-y)\log(1-\tilde{y})] \quad (1)$$

이미지 캡셔닝 모델에서 생성하는 캡션의 최대 길이는 20 으로 고정하였다. 캡션은 설명문 데이터에서 <BOS> (Begin of sentence) 토큰에서 시작하여 <EOS> (End of sentence) 토큰 전까지 출력된 단어들을 생성한 캡션이라고 간주하였다.

행동 인식 모델의 경우 UCF-101 데이터 셋[11]으로 선행 학습한 모델을 사용하였다. 행동 인식 모델은 Inception V3 모델에 비선형성 증가와 분류를 위하여, Dense 레이어 2 개를 추가하여 사용하였다. 이때 2 개의 Dense 레이어 사이에 오버 피팅 방지를 위하여 드롭 아웃을 추가하였고, 0.5 계수를 주었다. 학습을 진행할 때 파인 튜닝 방법을 사용하였다. 상위 레이어를 학습시킬 때 rmsprop 옵티마이저를 사용하였다. 분류기를 학습시킬 때 SGD 옵티마이저를 사용하였으며, 모멘텀은 실험적으로 가장 효율이 좋은 0.9, 러닝레이트는 초기 0.001 로 시작하여, 손실 값의 변동이 없을 경우 줄여 나가는 방식의 학습을 진행하였다.

3.3 데이터 셋

3.3.1 이미지 캡셔닝 데이터 셋

이미지 캡셔닝 데이터 셋은 Flickr 8K 를 사용하였다. Flickr 8K 는 8000 여개의 이미지와 각 이미지마다 5 개의 설명문을 가지고 있는 데이터 셋이다. 데이터 셋의 설명문을 임베딩 레이어에 입력하기 위하여 설명문의 문장들을 단어 단위로 쪼갠 후 단어의 출현 빈도 수를 key, 단어를 value 로 하는 딕셔너리를 만든다. 생성된 딕셔너리를 사용하여 입력하는 각 문장들을 인덱싱하여 1 차원 벡터로 만든 후 임베딩 레이어에 입력하며 학습한다.

3.3.2 행동 인식 데이터 셋

행동 인식 데이터 셋은 UCF-101 데이터 셋을 사용하였다. UCF-101 데이터 셋은 University of Central Florida 에서 제작한 인간 행동 인식 모델로, 101 가지의 클래스가 입력되어 있다. UCF-101 데이터 셋의 클래스는 평가하고자 하는 Flickr 8k 데이터 셋의 설명문과 문법이 다르기 때문에 어색하지 않은 문장 생성을 위하여, 본 논문에서 UCF-101 데이터 셋을 이미지 캡셔닝 모델의 데이터 셋에 포함된 설명문과 문법을 비슷하게, 소문자로 바꾸기, 띄어쓰기 등의 수정을 하였다. 수정한 결과 49 가지의 클래스가 남았고, 학습 모니터링을 위하여 학습 데이터셋과 동일한 클래스를 갖도록 테스트, 검증 셋을 수정하였다.

4. 실험 및 분석

실험을 위하여 Ubuntu 18.04 LTS 환경에서 Python 딥 러닝 라이브러리 중 하나인 Keras 를 사용하여 본 논문에서 제안한 모델을 구현하였다. 평가는 Flickr 8K 데이터 셋의 테스트 셋을 사용하였다. 테스트 셋은 총 1000 장으로 구성되어 있다. 실험 결과는 그림 3 과 같다. 그림 3 에서 im2txt 는 기존의 이미지 캡셔닝이며, 제안하는 모델로 생성된 캡션에서 빨간 부분으로 기입된 단어들은 행동 인식 모델로 생성된 동사 부분 단어들이다. 실험 결과로 보아, 제안하는 방법을 사용하면, 기존의 이미지 캡셔닝 보다 좀 더 행동 인식에 강한 캡션을 출력할 수 있다. 그러나, 행동 인식 모델의 문법이 평가 대상인 문법과 다소 다른 점이 있어, 기존의 이미지 캡셔닝 평가 방법인 BLEU, CIDEr 과 같은 방법을 이용할 수 없었다. 또한 행동 인식 모델의 데이터 셋이 UCF-101 인 관계로 인간의 행동만 판단할 수 있다. 그러나, Flickr 8K 데이터 셋은 각 종 동물들의 이미지도 입력되어 있기 때문에 인간이 주체가 아닌 이미지는 상대적으로 정확도가 낮다.

5. 결론 및 향후 연구

본 논문에서 행동 인식 모델을 추가하여 참조하는 이미지 캡셔닝 모델을 제안 및 실험하여 기존의 이미지 캡셔닝 보다 행동을 추정하는데 특화된 것을 보였다. 또한 여러 모델들 (e.g. 행동 인식, 표정 인식, etc.)의 결과 값을 자연어 처리 모듈로 합쳐 새로운 캡션을 만들어낼 수 있음을 보였다. 본 논문에서 제안하는 방식을 이용하면 높은 정확도의 형식이 정해져 있는 캡션을 생성할 수 있을 것이다.

Acknowledgement

본 연구는 경기도의 경기도 지역협력연구센터 사업의 일환으로 수행하였음. [GRRC-가천 2017(B01), 시니어 라이프 로그 기반 행동 분석]

참고문헌

[1] Alex Krizhevsky, et al. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (NIPS), 2012

[2] Hodosh, et al. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, 47, pp. 853-899,

[3] Kyunghyun Cho, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Empirical Methods in Natural Language Processing (EMNLP), 2014

[4] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation, 1997, 9, 8, pp. 1735-1780

[5] Oriol Vinyals, et al. Dumitru Erhan; Show and Tell: A Neural Image Caption Generator. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164

[6] Kelvin Xu, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. International Conference on Machine Learning (ICML), 2015

[7] Justin Johnson, et al. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4565-4574

[8] Karen Simonyan, Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems (NIPS), 2014, 27

[9] Diogo C. Luvizon, et al. 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5137-5146

[10] Christian Szegedy, et al. Rethinking the Inception Architecture for Computer Vision. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826

[11] Khurram Soomro, et al. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. ArXiv, 2012