인물 정보 태그를 활용한 드라마 영상의 이미지 캡셔닝 성능 개선

김현주 ○, 조진욱, 현지웅, 정윤경

성균관대학교

{julia1028, brbl, kabbi159, aimecca} @ skku.edu

Quality Improvement of Image Captions for Cinematic Video Using Character Tags

HyunJu Kim○, JinUk Cho, Jiwung Hyun, YunGyung Cheong

SungKyunKwan University

요 약

이미지 캡션 생성 (image captioning task) 은 주어진 이미지를 설명하는 자연어 문장을 생성하는 과제이다. 최근 딥러닝 기술을 기반으로 하는 이미지 캡션 모델들은 이미지를 설명하는 능력이 준수하다. 그러나 이런 범용적인 모델들이 생성한 캡션은 등장인물의 이름 등 고유한 정보가 포함되지 않는다. 특히, 드라마나 영화와 같은 분야에서는 고유 정보가 포함되지 않을 경우 캡션의 설명력이 부족하다. 본 논문에서는 기존 높은 성능의 방법론 중 학습 시 이미지와 객체 정보를 연결해주는 객체 태그(object tag)를 도입한 OSCAR+ (VinVL[2]) 모델을 기반으로, 객체 태그에 등장 인물의 태그(character tag)를 추가하여 객체 정보 뿐만 아니라 등장 인물의 정보까지 이미지와 연결되도록 한다. 한국 드라마 "미생"에서 수집한 데이터로 OSCAR+ [2] 모델을 미세 학습 (fine-tuning)한 결과, 적은 양의 데이터로도 등장 인물 이름과 같은 추가 정보를 포함하는 캡션을 생성하도록 학습하는 것이 가능함을 확인하였다.

1. 서 론

최근, 시각 정보와 언어 정보를 동시에 학습하여 모델이 좀 더 다양한 정보를 가지고 지식을 구축하도록 하는 시도가 많이 일어나고 있다. 그러한 시도들을 Visual Language tasks라고 하는데, Visual Question Answering, Visual Commonsense Reasoning, Image-Text retrieval, Image Captioning 등 다양한 세부 task가 있다.

그 중 이미지 캡셔닝(Image Captioning)이란, 이미지가 보여주고 있는 것이 무엇인지 그 시각적 정보를 해석하여 이미지에 대한 설명, 즉 캡션을 자연어로 생성하는 것을 말한다. [1], [2], [3], [4], [5] 등 많은 논문들이 이미지 캡셔닝을 포함한 Visual Language tasks를 해결하고자 노력하는데, 그 중 OSCAR[1]와 VinVL[2]은 Transformer 모델을 기반으로 한 구조를 제안한다. 그들은 이미지에서부터 특성 벡터와 객체 태그를 추출하며, 해당 이미지와 연관된 자연어 문장을 붙여 세가지 종류로 이루어진 데이터를 구성하고 multi-layer Transformer에 넣어 이미지와 자연어의 cross-modal representation을 학습시킨다. VinVL의 경우 OSCAR가 사용한 이미지 추출 방식과 손실 함수를 개선시킨 OSCAR+ 모델을 제시하였으나, 객체 태그를 사용하는 구조는 동일하다. 두 모델 모두 객체 태그 덕분에 모델은 그것을 활용하지 않은 것에 비해 이미지와 자연어 문장을 연결하는 과정을 더 수월하게 학습하여 좋은 성능을 기록하였다.

General-purposed Model (OSCAR+) : a woman standing next to a man in a suit
expected caption : sangsik and jiyoung are in a conversation

그림 1 범용적 모델 결과와 드라마 인물 정보를 담은 캡션

드라마나 영화와 같이 영상을 분석하고자 할 때, 이런 이미지 캡셔닝 기술은 많은 도움이 될 수 있다. 예를 들어, 그 영상의 줄거리를 요약할 때 시각적인 정보와 대본이나 장면 설명문과 같은 언어적인 정보를 모두 담은 지식을 활용하는 것이 더 좋은 질의 내용을 담을 수 있다. 또한, 영상을 저차원의 데이터로 변환할 수 있다는 것 역시 다양한 시도를 하는데 있어 많은 도움이 된다. 그러나 범용적인 목적으로 학습된 이미지 캡셔닝 모델은 드라마와 영화 분야 고유의

* 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원과 (No.2019-0-00421, 인공지능 대학원지원) 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2017-0-01642)

지식은 학습하지 않은 상태이다. 특히, 하나의 드라마와 영화를 설명하거나 이해하는데 빠질 수 없는 등장인물에 대한 지식은 범용적인 모델에서는 얻을 수 없는 지식이다. ([그림 1] 참고)

따라서, 이번 논문에서는 드라마와 영화 분야의 영상 데이터를 그 분야에 맞게 처리하여 OSCAR와 VinVL에 기술된 이미지 캡셔닝 모델의 구조를 참고 및 변형해 필요한 기반 지식이 잘 학습되는 것을 목표로 한다. OSCAR와 VinVL(OSCAR+)에서는 객체 태그만을 사용하였지만, 인물 태그를 추가해 시각적 정보와 그 드라마 혹은 영화의 등장인물을 연결시키기 쉽도록 하여 그 고유의 정보가 학습될 수 있도록 하였다.

그 결과, 본 논문이 기여할 바는 다음과 같다. 1) 먼저 드라마나 영화와 같이 시각적 정보 외에 다른 정보가 있어 이미지만으로는 캡션을 생성하기 어려운 분야에서 더 나은 캡션을 만들어낼 수 있다. 2) 다음으로 인물 태그를 추가하는 간단한 방식으로도 인물 정보를 캡션에 포함할 수 있도록 해 응용하기 쉽다.

2. 접근 방법 (Approach)

2.1 모델 (Model)

드라마나 영화 분야의 경우 더 나은 지식 활용을 위해 자체 정보가 포함된 캡션이 필요하다. 따라서, OSCAR+[2] 모델 구조를 기반으로 새로운 태그를 추가하여 모델이 생성할 캡션이 드라마나 영화 자체의 정보를 담을 수 있도록 하였다. 특히, 본 논문에서는 이미지만으로는 추출해내기 어렵지만 드라마나 영화 분야에서 무시할 수 없는 등장 인물의 이름에 초점을 맞추고 있다.

2.2 인물 태그 (Character tag)

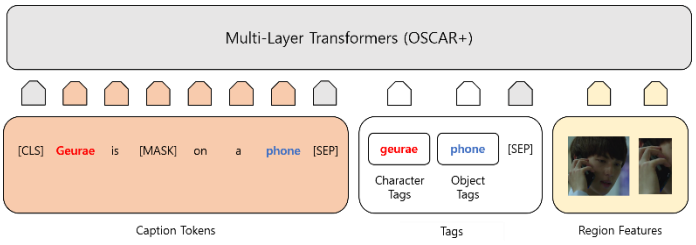

그림 2 미세 학습에 사용한 모델 구조 (인물 태그 추가)

OSCAR+ 모델은 단어-태그-이미지의 3-튜플 (w, q_o, v) 형태의 데이터를 입력으로 받는다. 여기에서 w 는 캡션 내 단어들의 임베딩 시퀀스(sequence of word embedding)이고, q_o 는 이미지에서 감지된 객체의 단어 임베딩인 객체 태그(object tag), 그리고 v는 이미지의 특정 영역에서 추출된 특성 벡터(regional feature vectors)를 의미한다.

이미지 캡셔닝을 위해 드라마 데이터셋으로 미세 학습(fine-tuning)을 진행할 때 OSCAR+ 모델의 객체 태그(object tag) 이외에도 인물 태그(character tag) q_c 를 [그림 2]와 같이 추가하면 모델은 시각적인 정보나 언어적인 정보 외에도 드라마 내 정보를 받을 수 있게 된다. 여기서 q_c는 v가 추출된 이미지에 등장하는 인물의 이름에 해당하는 단어 임베딩이며, 따라서 최종 입력 형태는 (w, q_o, q_c, v)의 4-튜플 형태로 주어진다. 이 때 q_c와 q_o의 데이터 형태가 같기 때문에, 큰 노력 없이 데이터셋에 추가가 가능하며 쉽게 학습에 사용할 수 있다.

3. 실험 (Experiment)

3.1 데이터 (Dataset)

미세 학습을 하기 위한 데이터로 한국 드라마 "미생"의 1화부터 10화까지의 영상을 활용해 구축한 데이터[6]를 활용한다. 인물의 행동을 기준으로 영상을 짧은 시간 단위로 나누고 이미지를 균일하게 샘플링 하였다. 이후, 샘플링된 이미지에 등장하는 인물을 인물 태그로 변환하고, 그 이미지에 대응되는 캡션으로는 사람이 작성한 것을 사용하였다. 이때 인물 태그는 드라마 “미생”의 등장인물 이름으로 표현하였으나, 주요 인물과 달리 자주 나오지 않거나 이름이 나오지 않는 경우는 그들을 특정할 수 있는 ‘interns’나 ‘executives’와 같이 태깅하였다. 또한, 각 이미지의 객체 태그와 특성 벡터는 VinVL과 동일한 방식으로 추출하였다.

3.2 세부 구현 사항 (Implementation Detail)

본 논문에서는 BERT base 모델의 가중치로 초기화된 $OSCAR+_B$ 모델 구조를 사용해 미세 학습(fine-tuning)을 하였다. 이 모델은 64의 batch size와 3e-5의 학습률(learning rate)로 60 에폭 동안 사전 학습된 가중치를 가진다.

미생 데이터셋[6]으로부터 생성한 데이터 약 6000개를 8:2 비율로 나눠 각각 학습과 검증에 사용했다. 미세 학습 시 $OSCAR+_B$ 모델의 초매개변수(hyperparameter) 값을 그대로 사용하되 레이블 평활화(label smoothing)를 0.2로 주었다. 추론 시 빔 검색(beam search)을 (n = 5)로 시행하였고 최대 생성 길이(max generation length)를 20으로 주었다.

4. 결과 (Results)

4.1 정량적 평가 (Quantitative Evaluation)

	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	METHOR	$ROUGE_L$
$Model_{q_o}$	51.42	41.54	35.39	30.38	26.81	50.48
$Model_{q_o+q_c}$	57.74	47.62	40.99	35.70	30.34	56.67
Δ	6.32 ↑	6.08 ↑	5.60 ↑	5.32 ↑	3.53 ↑	6.19 ↑

표 1 인물 태그의 유무에 따른 모델 성능 비교

[표 1]은 미생 데이터셋[6]으로 미세 학습(fine-tuning)한 두 모델의 BLEU, METEOR, ROUGE_L score를 비교한 결과이다. 첫번째 모델 $Model_{q_o}$ 은 인물 태그 없이 기존 OSCAR+ 구조를 그대로 사용하여 학습 및 추론한 결과이고, 두번째 모델 $Model_{q_o+q_c}$ 은 인물 태그를 추가하여 학습 및 추론한 결과이다. 표에 나온 값에서 알 수 있듯이, 두번째

모델이 기록한 BLEU-score 35.7%의 기록이 첫번째 모델의 30.4%보다 약 5% 높은 것을 보아 ([표 1]의 세번째 행 참고) 인물 태그를 추가하여 학습한 것이 더 좋은 성능을 보이는 것을 알 수 있다.

4.2 정성적 평가 (Qualitative Evaluation)

그림 3 인물 태그를 사용한 모델 추론 결과 예시

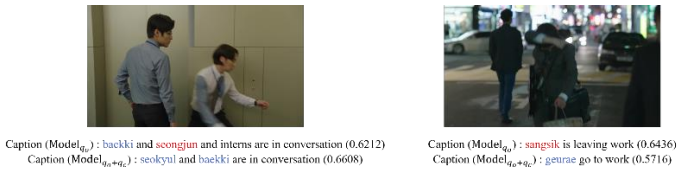

그림 4 인물 태그 유무에 따른 모델 추론 결과 비교

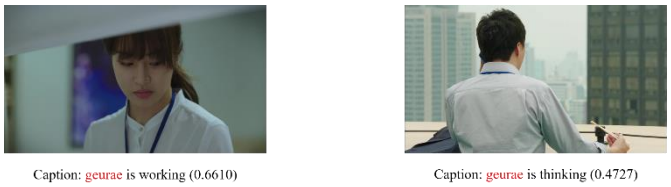

그림 5 모델의 잘못된 추론 예시

드라마라는 매체의 특성 상 데이터의 수가 많지 않고 또한 관련 연구가 적어 비교 대상을 선정하기 어려운 문제가 있다. 따라서 정량적 평가 외에도 정성적 평가를 통해 모델을 평가하고자 한다. ([그림 3, 4, 5] 참고. 알맞게 생성된 인물 이름은 파란색으로, 잘못된 이름은 빨간색으로 표기하였다.)

먼저, [그림 3]은 모델에 입력으로 주어진 이미지와 모델이 생성한 캡션 결과를 나타낸다. 위 그림을 보면 모델이 생성한 캡션은 이미지에 등장하는 인물과 상황을 잘 인지하는 것으로 확인된다.

[그림 4]는 기존 OSCAR+와 같은 형태의 모델과 인물 태그를 사용한 모델 간의 결과 비교이다. 두 모델 모두 이미지에 나타나는 상황과 인물의 행동은 잘 인지하지만 인물 태그를 사용한 모델이 등장인물의 이름을 잘 파악하는 경향을 보인다. 이로서 인물 태그를 추가하는 것이 모델로 하여금 인물 정보와 이미지 간 관련성을 더 잘 찾아내도록 이끈다고 볼 수 있다.

[그림 5]는 모델의 추론 결과가 정답 캡션과 상이한 것들을 보여준다. 이를 보면 인물을 잘못 추론한 경우 주인공 인물의 이름으로 치우치는 결과를 보인다. 드라마 특성상 주인공인 “장그래”가 등장하는 경우가 많아 데이터셋 편향으로부터 생겨나는 문제를 주된 이유로 고려할 수 있다.

5. 결 론

앞서 보인 것처럼 본 논문에서는 이미지 캡셔닝 기술을 드라마나 영화와 같이 이미지 외 추가 정보가 필요한 분야에 적용하기 위해, OSCAR+에 새로 인물 태그를 추가한 모델 구조를 제안했다. 또한 한국 드라마 "미생"으로부터 구축한 데이터셋으로 위 모델을 미세 학습함으로써 모델로 하여금 인물 정보가 포함된 캡션을 생성할 수 있음을 확인했다. 마지막으로, 논문에서 제안하는 모델이 객체 태그만 사용하는 기존의 OSCAR+ 모델 구조보다 드라마 내 정보를 담는 것에 있어서 더 나은 성능을 보임을 확인했다. 논문에서 제안하는 모델 구조를 통해 드라마나 영화 등 이미지만으로 설명하기 어려운 영상을 좀 더 다루기 쉽게 만들어 다양한 연구에 활용할 수 있기를 기대한다.

참 고 문 헌

[1] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." European Conference on Computer Vision. Springer, Cham, 2020.

[2] Zhang, Pengchuan, et al. "VinVL: Making Visual Representations Matter in Vision-Language Models." arXiv preprint arXiv:2101.00529 (2021).

[3] Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." arXiv preprint arXiv:1908.02265 (2019).

[4] Su, Weijie, et al. "Vl-bert: Pre-training of generic visual-linguistic representations." arXiv preprint arXiv:1908.08530 (2019).

[5] Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[6] 황지수, and 김인철. "AnoVid: 비디오 주석을 위한 심층 신경망 기반의 도구." 멀티미디어학회논문지 23.8 (2020): 986-1005.