자연어 기반 정성 평가 캡셔닝 생성 및 화질 정량 평가

CUAI 7기 MM1팀
김소원(소프트웨어학부), 나상현(소프트웨어학부), 나영은(역사학과), 최형용(국어국문학과)

[요약**]** 사진 입력에 대해서 정량적 인지 화질 점수와 자연어 기반 정성 평가 캡셔닝을 생성하는 AI 모델을 개발을 목표로 한다. “CPTR: Full Transformer Network for Image Captioning” 논문의 모델에 따라 코드 구조를 완성하고 실행 중 논문과 코드 간의 불일치하는 부분을 확인한다. 이를 활용하여 유의미한 결과를 도출한다.

1. 서 론

사용자의 상황에 최적화된 서비스를 제공하는 스마트폰 카메라의 AI 영상 처리 기능 개발 연구가 계속되고 있다. 최근 연구에서는 이미지 화질 평가에 대한 다양한 딥러닝 기반 접근과 최신 기술들을 다루고, 트랜스포머 기반 이미지 캡셔닝 기술을 통해 이미지와 자연어 처리의 융합으로 성능을 개선하는 연구에 집중하고 있다. 화질 평가에 대한 기준이 다양하기 때문에, 선명도, 노이즈 정도, 색감, 선호도 등 여러 인지적 화질 요소를 종합적으로 고려하여 정량 평가 점수를 예측하는 과정이 필요하다. 또한 단일 점수에서 누락될 수 있는 의미들을 자연어로 영상의 화질을 설명하여 기능을 개선할 수 있다.

본 연구는 카메라로 촬영된 영상의 화질에 대한 정량 평가 점수를 예측하고, 그 평가 결과를 자연어로 상세하게 표현하는 알고리즘을 개발하는 것을 목표로 한다.

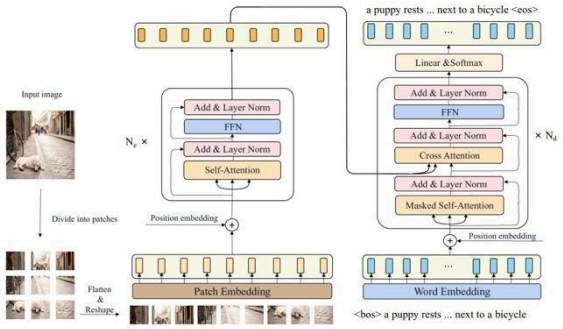
2. 본 론

1) 활용 데이터

본 연구는 데이콘 대회 ‘2023 Samsung AI Challenge: Image Quality Assessment”에서 주어진 데이터를 활용하였다.

train[폴더]에 학습용 input 이미지 파일(jpg) 54662장, test[폴더]에 추론용 Input 이미지 파일(jpg) 13012장, train.csv에 img_name(파일명), img_path(이미지 경로), mos(화질 평가 점수(0~10, float)), comments(화질 평가 캡셔닝 정보(text, 영어)), test.csv[파일]에 img_name, img_path로 구성된다.

2) 이론적 배경



<그림1> CPTR의 전반적인 구조

이 논문은 새로운 sequence to. sequence 예측 관점에서의 이미지 캡셔닝을 다룬다. CaPtion TransformeR (CPTR)는 sequentialized 원본 이미지를 Transformer의 입력으로 사용하는 모델로, 모든 인코더 레이어에서 처음부터 global context를 모델링할 수 있으며, convolution-free 모델이라는 점을 강조한다.

피처를 CNN에서 추출하는 기존의 캡셔닝 모델과 다르게, CPTR은 원본 이미지를 sequentialize 해 입력 데이터로 바로 사용한다. 인코더는 self-attention 매커니즘을 통해 패치들 간의 장기 의존성을 처음부터 활용할 수 있게 하고, 이미지를 고정된 크기(16×16)로 작은 패치로 나눈 후, 각 패치를 flatten하고 이를 1D 패치 시퀀스로 변형하여 활용한다. 디코더의 “words-to-patches” attention은 정확하게 해당 패치에 집중해 단어를 예측할 수 있게 한다. MS COCO 데이터셋을 활용하여 "CNN+Transformer" 모델보다 성능이 향상된 것을 확인한다.

3) 코드 구조

Processing, Learning, Describe 3단계를 거쳐서 코드 구조를 완성했다.

Processing 단계에서 캡션 데이터 어휘집 추출, 어휘집 기반으로 토큰화 처리, 이미지에서 피처 추출을 한다.

Learning 단계에서 트랜스포머 모델 적용(체크포인트가 있다면 로드)하고 최적화 옵티마이저(Adam)와 손실함수(CrossEntropyLoss) 설정한다. (div by 0 등의 에러가 없도록) NaN이나 inf 등의 값을 처리하고 기울기 폭발 문제 방지를 위한 gradient clipping, 학습 종료시 체크포인트 저장 및 학습 종료 로그 기록한다.

Describe 단계에서 어휘집 로딩, resnet152으로 테스트 데이터의 이미지 피처 추출, Beam search 활용 캡션 생성, CLIP ViT 활용하여 랭킹, 최고득점 캡션을 csv에 저장한다.

4) 논문과 코드 간 불일치

CPTR은 Full transformer 이고 convolution-free 라는

논문의 내용과 달리, 코드에서는 이미지 피쳐를 pre-trained CNN (ResNet 152)를 사용해 추출해 완전한 convolution-free 모델은 아닌 것을 확인하였다. 코드에서 CNN(resnet) + Transformer encoder를 쓰는 것을 확인했고 ViT 모델을 랭커로만 활용한 것을 확인하였다.

5) 결과
데이콘에 베이스라인 모델의 결과와, 본문에서 소개한 모델의 결과를 제출하여 비교하려 하였으나 베이스라인 모델의 캡션 생성 오류로 진행하지 못하였다. 다만 본문 모델 캡션의 결과는 0.66175점으로, 캡셔닝 점수 기준 23등의 성과를 거두었다.

6) 연구 한계 및 제언
베이스라인 모델로 추론한 mos와 캡션을 데이콘에 제출하지 못하여 베이스라인 모델에 비한 본문 모델의 향상 정도를 수치적으로 나타내기 어려웠다.
CNN으로 이미지 피처를 추출한 후 이를 트랜스포머 인코더에 입력으로 주는 방식을 적용하여 성능이 기대만큼 향상되지 않았으리라 생각한다. 향후 CLIP 등의 방대한 양의 데이터로 pre-trained ViT 인코더를 사용하여 성능 향상을 기대할 수 있을 것이다.

3. 결 론

데이콘에서 주어진 이미지 데이터를 기반으로 MOS를 예측하고 캡션을 생성하였다. 이를 통해 트랜스포머 구조 기반 모델을 사용한 이미지 캡셔닝을 수행할 수 있었다.
따라서 자연어 처리에서 효과적인 성능을 보이는 트랜스포머 모델에 완전한 convolution-free 구조를 적용하기 위해 추후 pre-trained 인코더 등의 모델을 활용하는 연구를 지속하면 이미지 캡셔닝을 효과적으로 수행하는 모델을 개발할 수 있을 것으로 확인된다.

참고 문헌
1) Wei Liu, et al. “CPTR: Full Transformer Network for Image Captioning”, arXiv:2101.10804, 2021.
2) Tsung-Yi Lin, et al. “Microsoft COCO: Common Objects in Context”, arXiv:1405.0312, 2015
3) Ashish Vaswani, et al. “Attention Is All You Need.”, Google AI Language, 2017