ChatGPT 기반 한국어 Vision-Language Pre-training을 위한 고품질 멀티모달 데이터셋 구축 방법론

성진*1, 한승헌*2, 신종훈1, 임수종1, 권오욱1
1 한국전자통신연구원, 2 충북대학교
{real_castle, jhshin82, isj, ohwoog}@etri.re.kr, hhanheon@gmail.com

High-Quality Multimodal Dataset Construction Methodology for ChatGPT-Based Korean Vision-Language Pre-training

Jin Seong*1, Seung-heon Han*2, Jong-hun Shin1, Soo-jong Lim1, Oh-woog Kwon1
1 Electronics and Telecommunications Research institute (ETRI)
2 Chungbuk National University

요 약

본 연구는 한국어 Vision-Language Pre-training 모델 학습을 위한 대규모 시각-언어 멀티모달 데이터셋 구축에 대한 필요성을 연구한다. 현재, 한국어 시각-언어 멀티모달 데이터셋은 부족하며, 양질의 데이터 획득이 어려운 상황이다. 따라서, 본 연구에서는 기계 번역을 활용하여 외국어(영문) 시각-언어 데이터를 한국어로 번역하고 이를 기반으로 생성형 AI를 활용한 데이터셋 구축 방법론을 제안한다. 우리는 다양한 캡션 생성 방법 중, ChatGPT를 활용하여 자연스럽고 고품질의 한국어 캡션을 자동으로 생성하기 위한 새로운 방법을 제안한다. 이를 통해 기존의 기계 번역 방법보다 더 나은 캡션 품질을 보장할 수 있으며, 여러가지 번역 결과를 앙상블하여 멀티모달 데이터셋을 효과적으로 구축하는데 활용한다. 뿐만 아니라, 본 연구에서는 의미론적 유사도 기반 평가 방식인 캡션 투영 일치도(Caption Projection Consistency)를 소개하고, 다양한 번역 시스템 간의 영-한 캡션 투영 성능을 비교하며 이를 평가하는 기준을 제시한다. 최종적으로, 본 연구는 ChatGPT를 이용한 한국어 멀티모달 이미지-텍스트 멀티모달 데이터셋 구축을 위한 새로운 방법론을 제시하며, 대표적인 기계 번역기들보다 우수한 영한 캡션 투영 성능을 증명한다. 이를 통해, 우리의 연구는 부족한 High-Quality 한국어 데이터 셋을 자동으로 대량 구축할 수 있는 방향을 보여주며, 이 방법을 통해 딥러닝 기반 한국어 Vision-Language Pre-training 모델의 성능 향상에 기여할 것으로 기대한다.

주제어: ChatGPT, 멀티모달, 데이터 생성, 이미지 캡션, 문장 유사도,

1. 서론

멀티모달 데이터셋은 이미지 캡셔닝 및 시각-언어 모델 학습과 같은 다양한 컴퓨터 비전 및 자연어 처리 태스크에 중요한 역할을 한다. 특히, 최근 몇 년간 이미지와 텍스트를 결합한 모델의 발전으로 멀티모달 데이터셋의 필요성이 더욱 커졌다. 그러나 한국어로 된 대규모 멀티모달 데이터셋의 부족으로 인해 이러한 모델의 학습 및 평가가 제한되어 있다.

이러한 상황에서, 우리는 한국어 Vision-Language Pre-training(VLP)[1]모델을 학습하기 위해 대규모 시각-언어 멀티모달 데이터셋을 구축하는 방법을 연구하고자 한다. 현재, 한국어로 된 시각-언어 데이터셋은 부족하며, 양질의 데이터 획득이 어렵다. 따라서, 이 연구에서는 기계 번역 기술을 활용하여 외국어(영문)로 작성된 시각-언어 데이터를 한국어로 번역하고, 이를 활용하여 대규모의 데이터셋을 생성하는 방법을 제안한다.

우리의 주요 기여는 ChatGPT를 활용하여 자연스럽고 고품질의 한국어 캡션을 자동으로 생성하는 새로운 방법을 제안한 것이다.

* 동일한 기여도

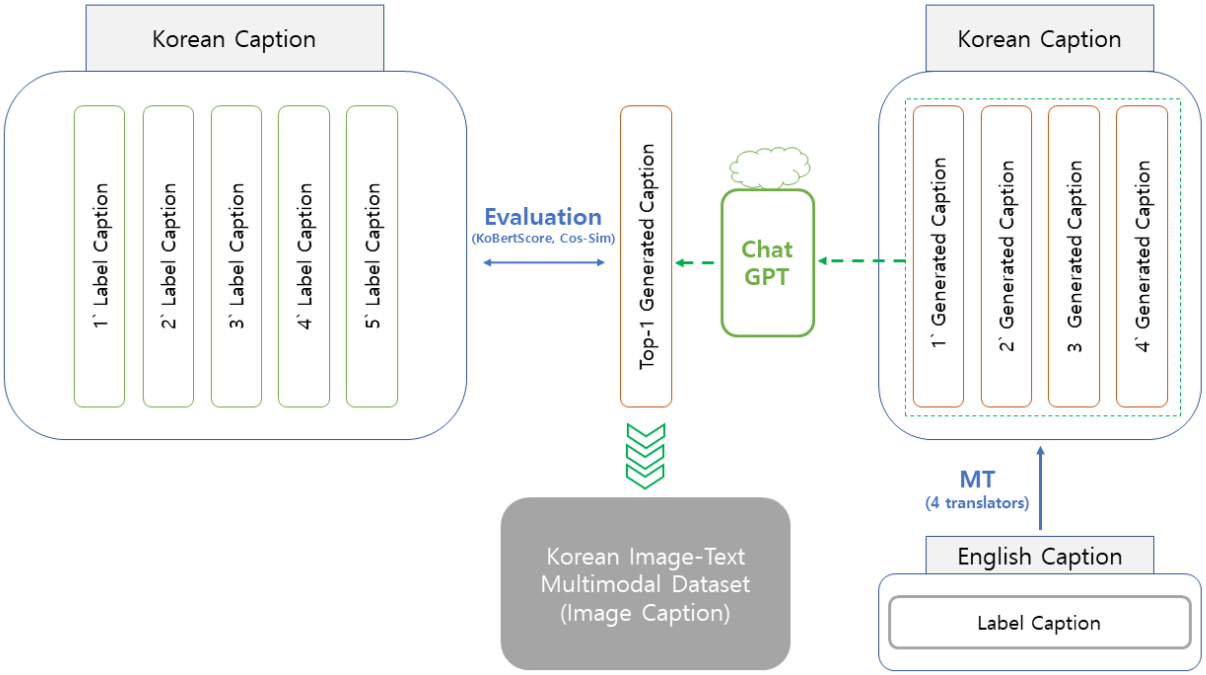

그림 1. ChatGPT 기반 멀티모달 이미지 캡션 데이터 생성 및 평가 시스템 구조도

이를 통해 우리는 기존의 기계 번역 방법보다 더 나은 캡션 품질을 보장할 수 있으며, 여러가지 번역 결과를 앙상블하여 멀티모달 데이터셋을 효과적으로 구축할 수 있다. 또한, 이 연구에서는 의미론적 유사도를 기반으로 한 평가 방식인 캡션 투영 일치도(Caption Projection Consistency)를 소개하고, 다양한 번역 시스템 간의 영-한 캡션 투영 성능을 비교하며 이를 평가하는 기준을 제시한다.

최종적으로, 본 연구는 대표적인 SOTA(State-of-the-Art) 기계 번역기들보다 우수한 영한 캡션 투영 성능을 입증한다. 이를 통해, 우리의 연구는 부족한 High-Quality 한국어 데이터셋을 자동으로 대량 구축할 수 있는 방향을 제시하며, 이 방법을 통해 딥러닝 기반 한국어 Vision-Language Pre-training 모델의 성능 향상에 기여할 것으로 기대한다.

주요 기여 정리

- 한국어 시각-언어 데이터 셋 구축시, High-Quality 한국어 캡션 생성 및 구축 아이디어를 ChatGPT와 기계번역기를 모두 활용하여 제안.
- 기계번역을 통해 외국어(영문) 시각-언어 데이터를 한국 데이터로 구축을 할 때, 의미론적 유사도 평가 기반의 캡션 투영 일치도(Caption Projection Consistency)평가가 필요하다. 이를 위한 평가 기준을 제안 및 가이드라인 제시.

2. 관련 연구

풍부한 영어 멀티모달 자원에 비해, 한국어 시각-언어 멀티모달 자원은 매우 부족한 상황이다. 이 문제를 해결하기 위해 기존 연구자들은 주로 두 가지 방법을 사용해왔다. 첫 번째 방법은 캡션 품질을 보장하기 위해 대규모의 자원이 필요한 Human Annotation을 활용하는 것이며, 두 번째 방법은 자원이 상대적으로 작더라도 번역 가능한 외국어로 된 이미지 캡션을 활용하는 방법이다.

2.1 Human Annotation

이미지 캡션 데이터를 구축하기 위해 일반적으로 Human Annotation방식을 사용하는데, 이는 작업자들을 활용하여 이미지에 대한 설명 또는 주석을 생성하는 방법이다. Human Annotation은 크게 Microtask Annotation, 크라우드소싱(crowd sourcing) 2가지 방식으로 나뉜다.

먼저, Microtask Annotation 방법은 이미지 캡션 주석을 얻을 때, 작업을 미세하게 나누어 더 쉽게 처리하는 아이디어이다. 예를 들어, 이미지에서 특정 객체나 장면에 대한 주석을 수집하는 작업을 상대적으로 세세하게 나누어 처리할 수 있다.

다음으로, 이미지 캡션 데이터를 구축하기 위해 많이 채택되는 방식인 크라우드소싱(crowdsourcing)이다. 크라우드 소싱은 온라인 플랫폼을 통해 대규모 작업자 집단, 즉 "크라우드"를 활용하여 다양한 작업을 수행하고 데이터를 수집하는 방법이다. 이를 통해 이미지 캡션 데이터를 수집하려면, 작업자들은 이미지에 대한 설명 또는 주석을 생성하거나 선택한다. 크라우드 소싱은 데이터 확보의 편의성과 비용 효율성 면에서 장점을 가지지만, 데이터 품질 관리가 어려우며, 작업자의 정확성과 노력에 따라 품질의 변동성이 발생할 수 있다. 또한 작업자들의 배경과 문화적 차이로 인해 데이터 편향성이 나타날 수 있으며, 개인 정보 보호와 명확한 작업 지시 사항 제공이 중요한 고려 사항이다. 이러한 장단점을 고

그림 2. 캡션 투영 성능 평가 데이터 셋(총 5000개 이미지-캡션 쌍 멀티모달 데이터)

려하여 크라우드 소싱을 효과적으로 활용하기 위해서는 데이터 품질 관리와 작업자 편향성 등을 주의 깊게 고려해야 한다.

일반적으로 Large Vision-Language Pretrain(VLP) Model을 사전학습시, Human Annotation 방법으로 구축한 시각-언어 데이터셋이 품질을 보장할 수 있으며, 데이터의 신뢰성이 높기 때문에 이 방식을 선호한다. 그러나, 이런 방법을 사용하는데 상당한 시간과 비용이 소요되며, 특히 VLP 모델의 학습을 위해 대규모 데이터셋을 수집하려면 더 많은 노력과 비용이 필요하다. 또한, 작업자들 간의 일관성을 유지하고 품질을 향상시키는 데에도 관리와 감독이 필요하며, 이로 인해 관리 부담도 피할 수 없는 부분이다.

2.2 기계 번역(Machine Translation)

Machine Translation을 사용한 연구들은 데이터 부족 문제를 비교적 저렴하며, 간단히 해결하기 위해 이미지 캡션 데이터를 영어에서 한국어로 번역하여 구축하는 방법을 채택한다.

기계 번역 방식을 활용하는데에는 여러 가지 번역 엔진이 존재하며, 대표적으로 파파고(Papago), Google, DeepL, ChatGPT 등이 있다. 이러한 번역 시스템은 각자 고유한 특징과 성능을 갖고 있습니다. 파파고(Papago)는 한국어 번역 분야에서 널리 사용되며, 실시간 번역 기능과 어휘력을 토대로 다양한 어휘 데이터를 활용하는 강점을 가지고 있다. Google 번역 엔진은 전 세계적으로 널리 사용되며, 다양한 언어 간 번역을 지원하는 데 강점을 지닌다. DeepL은 딥러닝을 기반으로 한 번역 엔진으로, 정확성과 자연스러움을 강조하며 전문 번역 분야에서 사용된다. 마지막으로, ChatGPT는 자연어 생성 모델로 사용되며, 번역 엔진으로써 사용되지는 않는다. 하지만, 딥러닝 기반으로 문장을 분석하여 사용자의 prompt에 맞는 내용을 출력해주는 특징을 가지며, 자기 회귀 모델(auto-regressive model) 특성에 맞춰 어색한 문장을 자연스럽게 표현할 수 있으며, 초거대 데이터 셋으로 학습이 되었기 때문에 노이즈에 강한 캡션 투영 능력을 기대할 수 있다.

그러나 이 방법은 몇 가지 한계를 가진다. 먼저, 기계 번역을 통한 캡션은 자연스럽지 않을 수 있으며, 번역 과정에서 문맥을 고려하지 않고 단어 단위로 번역되기 때문에 번역된 캡션은 자연스럽지 않을 수 있다. 또한, 데이터를 구축할 때 어떤 번역 시스템을 선택해야 하는지 결정을 내려야 하는데, 이때 번역 엔진 시스템 간의 캡션 투영 평가 기준이 명확하지 않다는 한계가 존재한다.

Machine Translation을 활용한 데이터 구축 방법은 비교적 저렴하고 신속히 데이터 확보가 가능한 장점이 있지만, 번역된 캡션의 자연스러움과 캡션 투영 성능 평가에 대한 어려움을 고려해야 한다.

2.3 의미론적 유사도 기반 평가 방법

KoBERTScore[3,4]**.** BERTScore는 생성된 텍스트 문장의 유사성을 자동 평가하기 위한 문맥 임베딩 기반 평가 방법으로, 단순히 문장의 표면적 형태를 고려하는 것이 아니라 문맥적 의미를 고려하여 유사성을 측정하는 지표로 알려져 있다. 이

와 유사한 방법으로 한국어 데이터를 활용하여 학습된 KoBERTScore가 존재하며, 이를 통해 한국어 문장에 대한 평가 정확도를 향상시킨 연구가 있다. 이를 통해 생성된 이미지 캡션의 문장을 정답 문장과 비교하여 의미론적 유사도(Semantic Similarity) 를 평가할 수 있다.

SentenceBERT(SBERT)[2]**.** SentenceBERT는 KoBERTScore과는 또 다른 문장의 의미론적 유사도 측정 방법이 가능한 BERT 기반 문장 이해 모델이다. 문장의 의미적 유사성을 평가하는 용도로 사용되는데, 주로 STS(Semantic Textual Similarity)와 같은 다운스트림 태스크를 해결하기 위해 사용되거나, 언어 모델의 문장 이해능력을 평가하기 위해 사용될 수 있다.

3. 연구 방법론

본 연구에서는 위의 문제들을 효율적으로 해결하기 위해 이미지-텍스트 멀티모달 캡션 구축시, 기본적으로 Machine Translation(MT)을 활용한다. 그러나, 단순히 기계번역을 활용하는 것이 아닌, 기계번역 통해 얻은 시각-이미지 캡션 데이터 결과를 ChatGPT의 Prompt로 활용하여 새로운 한국어 캡션을 생성하는 방법을 제안한다. 우리는 새로 제안한 방법인 GPT-4(generate)와 주요한 번역 시스템들간 '캡션화 투영 성능(Caption Projection Consistency)'을 평가 및 비교하기위해 의미론적 유사도 기반 평가 방법의 KoBERT score와 SBert의. 더불어, 생성형 언어모델을 활용하여 효과적이며 효율적인 새로운 앙상블 기반 데이터 구축 방법론을 제안한다. [그림 1]

3.1 모델 수행 과정

한글 캡션 생성을 위한 ChatGPT의 Prompt는 다음과 같다. 그림 1과 같이, 먼저 영문 캡션(Label Caption)을 서로 다른 4가지 SOTA 번역기의 입력으로 넣어 4가지 버전의 한국어 번역 캡션을 생성한다. 이렇게 생성된 4개의 번역 캡션을 ChatGPT의 입력으로 사용하는 동시, 최적의 한국어 캡션을 생성하기 위한 Prompt를 제시해준다. Prompt는 다음과 같다. "다음 번역문중 어떤 문장이 자연스러운지 선택(GPT4-select), 더 나아가 자연스러운 캡션을 생성하라(GPT4-generate)"

ChatGPT가 추론 과정을 거쳐, 4가지 번역문중 캡션 문장이 가장 자연스러우며 매치된 이미지를 잘 표현해주는 번역문을 선택한다(GPT4-select). 더 나아가, 4가지 캡션을 참조하여 새로운 자연스러운 캡션을 생성한다(GPT4-generate).

3.2 데이터셋 구축을 위한 프롬프트 엔지니어링

대량의 이미지-텍스트 데이터 셋을 구축하기 위해, ChatGPT 프롬프트 엔지니어링이 필수적이다. 이를 위해 처음엔, ChatGPT에 번역본과 함께 적합한 캡션으로 보이는 예시를 한 문장 정도 제공하였다(few shot). 하지만, 이 prompt를 사용할 경우, ChatGPT는 Top-1 Generated Caption을 출력할 때, 번역문 편향된 결과만을 출력하였다. (예시: 첫 번째 결과-Google, 두 번째 결과-Google, 세 번째 결과-Google … 1000번째 결과-Google)

이러한 편향된 출력의 문제를 해결하기 위해, ChatGPT의 [그림 3]과 같이 출력 형식을 Prompt를 통해 지정해준 뒤 수행해야한다. 우리는 출력 형식을 다음과 같이 설정했으며, 결과적으로 일관된 format 형태의 출력을 하는 동시, 편향되지 않은 결과를 보여주었다.

<Prompts for GPT4-Select >	<Prompts for GPT4-Generate>
the output format must be like {{ "Index" : , "translation" : (a) Top-1 Translator / (b) 선택된 캡션 }}	the output format must be like {{ "translation" : }} (c) 생성된 캡션

그림 3. 데이터 형식 지정을 위한 프롬프트

4. 실험 및 결과

4.1 평가 데이터 셋

한국어 이미지-텍스트 크로스 모달 평가용 데이터 셋.

제안된 실험에서 사용한 평가 데이터셋은 ETRI에서 구축한 "ETRI 이미지-텍스트 크로스 모달 평가용 데이터셋"으로, 데이터 구축시 Nocaps[6] 영문 캡션 데이터 셋에서 사용한 'Open Images'의 이미지를 동일하게 사용한다. 이 평가 데이터 셋은 총 15,100개의 이미지를 포함하며, 각 이미지당 5개의 서로 다른 캡션이 있다. 실험을 위해 데이터셋 중 1,000개의 이미지를 테스트를 위해 랜덤으로 Sampling을 하였으며, 추출된 테스트 셋을 사용하여 모델의 일반화 성능을 확인한다. 한국어 캡션은 한국어 문법 구조를 따르면서도 이미지의 핵심 의미를 정확하게 설명하도록 작성되었으며, 캡션 작성 시에는 이미지 내에 포함된 두 개 이상의 사물과 그들 간의 관계가 함께 고려되었다. 각 이미지에는 두 개 이상의 사물이 존재하며, 평균적으로 8개의 인스턴스가 이미지에 포함되어 있다. 이 평가 데이터셋은 'Nocaps' 영문 캡션 데이터셋을 참고하여 한국 및 영문 캡션을 모두 제공하므로, 영한 캡션 투영 성능 평가 데이터 셋으로 사용한다.

4.2 한국어 캡션 생성 문장 평가 방법

외형적 유사도 기반 평가(BLEU Score). 문장 유사도를 평가하는 전통적인 방법으로 BLEU(Bilingual Evaluation Understudy)[7]와 ROUGE(Recall-Oriented Understudy for Gisting Evaluation) [8]가 있다.

$$\mathbf{BP} = \begin{cases} \mathbf{1} & \mathbf{if\, c > r} \\ e^{\left(1-\frac{r}{c}\right)} & \mathbf{if\, c \le r} \end{cases}$$

$$\mathbf{BLEU} = \mathbf{BP} \times \exp\left(\sum_{n=1}^{N} \boldsymbol{w_n} \log \boldsymbol{p_n}\right)$$

위의 수식은 BLEU 평가 방법으로, N은 n-gram, p_n은 n-

gram을 위해 수정된 precision, w_n은 0~1사이 weight 값이며 n-gram 마다 weight이 주어지며 합은 1이 되어야 한

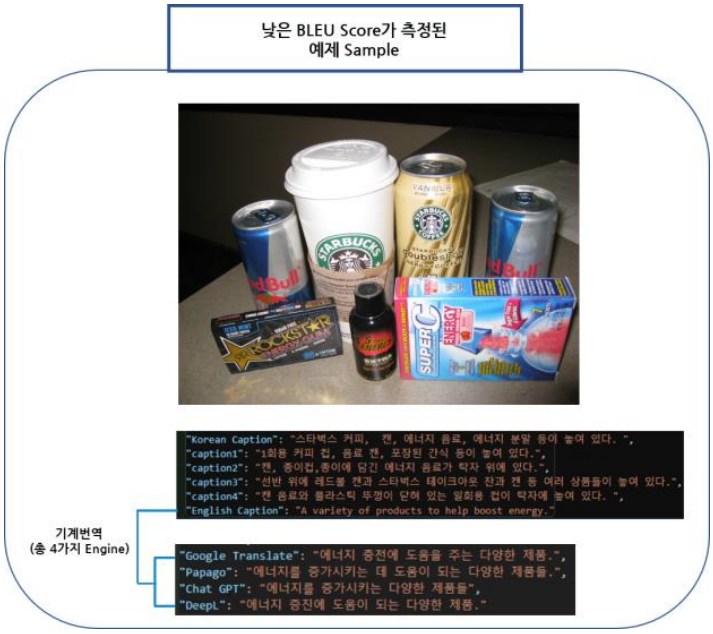

그림 4.BLEU Score가 낮게 측정되는 이미지-텍스트 쌍 예시

다. BP는 Brevity Penalty로, reference의 길이에 따른 penalty를 준다. c는 추론 문장의 길이, r은 추론 문장과 길이가 비슷한 값이다. 이 외형적 유사도 기반 평가 방법은 생성된 문장과 참조 문장 간의 토큰(token)수를 측정하며, 측정된 토큰들 간 n-gram 방식을 활용하여 통계적인 유사성을 평가한다. 그러나 이러한 방식은 문장의 외형적 유사성만을 고려하므로 의미가 동일한데도 불구하고 문장 내용이 다른 경우 평가시 낮은 점수를 부여할 수 있다.

의미론적 유사도 기반 평가(KoBERT Score, Semantic Score).

KoBERT Score[3,4]는 BERTscore[3] 기반 평가 방법으로, 한국어 참조 문장과 생성 문장을 각각 BERT의 입력으로 넣어, 문맥 벡터를 구한 후 각 토큰간의 코사인 유사도를 추론하여 문장간의 시멘틱 유사도를 측정할 수 있다.

Semantic Score도 KoBERT Score와 마찬가지로, 참조 문장과 생성 문장을 각각 SBERT[2]의 입력으로 넣어, 결과로 나온 문장 임베딩간의 유사도를 Pearson Correlation Coefficient를 사용하여 구한다.

실험은 한 이미지마다 5개의 서로 다른 한국어 참조 캡션이 달려있으며, 각 참조 캡션마다 생성 캡션사이의 유사도를 측정하여 총 5쌍의 캡션 투영 성능을 평균내어 평가한다. 생성된 캡션은 각 4개의 SOTA 번역기(DeepL, GPT3.5, Google, Papago)와 우리가 제안한 GPT4(Generate), GPT4(Select)이며, 각 생성된 캡션마다 투영 성능을 평가한다.

4.3 평가 결과

외형적 유사도 기반 평가 결과

표 1. BLEU Score로 측정한 번역기별 캡션 투영 성능 평가

Translator's Engine	BLEU Score
Papago	21.1
GPT3.5	14.7
Google	14.7
DeepL	15.7

표 1은 외형적 유사도 기반 평가 방식인 BLEU Score로 평가한 번역기별 캡션 투영 성능 평가 결과이다. Score가 대부분 14.7~21.1 사이의 값을 보이며 전반적으로 성능이 크게 떨어지는 모습을 보여준다. 그림 3의 정성적인 결과를 확인하였을 때, 참조 문장과 번역 문장간의 형태적으로 의미적으로 크게 다른 모습을 확인할 수 있다. 따라서, 우리는 캡션 투영 성능을 평가하기위해 전통적으로 사용되는 이미지 캡셔닝 평가 Method인 BLEU Score 대신, 딥러닝 기반 유사도 의미 기반 평가 방식인 KoBERTScore와 Pearson Correlation Coefficient을 사용한다.

BLEU score를 사용하여 각 번역 엔진별 캡션 투영 성능을 평가를 한 결과이다. 전반적으로 BLEU Score는 모든 번역 엔진에서 낮은 수치를 보인다. 그림 4은 이와 같은 .BLEU score 및 ROUGE score 모두 한계점을 가지는 외형적 유사도 기반 점수 평가 방식이기에, 원문 영어 캡션의 한영 번역한 결과가 이미지를 잘 묘사하는 캡션인지를 평가하기엔 적절하지 않은 Method라고 판단한다. 따라서 우리는 의미론적 유사도 기반의 평가 방식을 사용한다.

의미론적 유사도 기반 평가 결과

표 2. GPT기반 멀티모달 이미지-텍스트 데이터 캡션화 성능 평가 결과.

초록색 글씨:번역 엔진중 Best, **굵은 글씨**:전체 방법중 Best

Model	Method	Semantic Score	KoBERT Score
GPT4(Generate)	**Prompt**	**62.7**	68.2
GPT4(Select)	**w/Translator**	62.4	**68.4**
DeepL		62.3	68.2
GPT3.5	**Machine**	62.2	67.5
Google	**Translator**	62.1	67.5
Papago		61.8	68.0

표 2는 SOTA를 보이는 4개의 번역 엔진과, 우리가 제안한

GPT4 기반 2가지 한국어 캡션 구축 방식의 성능 비교 표이다. 평가 데이터는 "4.1 ETRI Caption Projection 성능 평가 데이터셋"을 사용하였으며, 총 2가지 의미론적 유사도 기반 평가를 하였다. 그 결과, 번역기들의 번역 문장과 캡션 생성을 위한 Prompt를 사용한 GPT4(Generate), GPT4(Select)가 Semantic Score, KoBERT Score 모두 기존의 번역기를 능가하며 SOTA를 보였다. 최종적으로, Semantic Score는 GPT4(Generate), GPT4(Select), 4가지 번역기(DeepL, GPT3.5, Google, Papago) 순으로 높은 성능을 보였으며, KoBERT Score는 GPT4(Select), GPT4(Generate), 4가지 번역기(DeepL, Papago, Google, GPT3.5) 순으로 높은 성능을 보였다. SOTA 번역기들의 캡션 투영 성능은 기존에 알려진 번역 성능과 동일하게 DeepL Semantic score 62.3, KoBERT Score 68.2로 가장 우수한 모습을 보인다.

한 가지 놀라운 점은 GPT3.5도 영한 캡션 투영 성능이 기존 모델들과 겨룰만한 정도로 좋다는 결과이다. 이 모델은 다른 기계 번역기들과 다르게 번역에 특화된 모델은 아니지만, 다양한 분야에 대한 학습을 바탕으로 번역 및 캡션에 대한 in-context learning(ICL)을 수행했기에 가능 한 것으로 보인다. 그러나, GPT3.5는 텍스트로만 학습을 수행했기 때문에, 사진과 같은 시각과 관련된 능력이 부족하다. GPT4는 시각-언어를 모두 이해하는 멀티모달 이므로, 주어진 번역기의 영문 캡션 활용하여 한국어 캡션 데이터를 구축할 때 캡션 투영 성능이 좋게 평가된 것으로 분석된다.

4.3 성능 분석 및 원인

이 연구 결과에 따르면, 현존하는 SOTA(State-of-the-Art) 영한 번역기 중, 가장 좋은 성능을 보여준 DeepL을 뛰어넘는 캡션 투영 성능을 보여준다. 특히, 단일 영문 캡션을 여러 가지 영한 번역기를 통해 한국어 캡션으로 생성한 후, 생성된 한국어 캡션을 GPT4를 활용하여 선택(GPT4-Select) 또는 새로 생성(GPT4-Generate)한 경우, 생성된 캡션은 이미지를 더 자연스럽게 표현하며 캡션의 품질이 향상되는 것을 확인할 수 있다.

이러한 결과를 바탕으로 우리는 대량 한국어 멀티모달 데이터셋 구축시, High-Quality 데이터 셋을 자동으로 생성할 수 있으며 이미지의 의미론적 유사도를 최대한 살려주는 데이터 셋을 구축할 수 있는 방법을 증명하였다. 또한, 이 데이터셋을 활용하여 멀티모달 모델 (예: Vision-Language Pre-training 모델)을 학습할 때, 번역 말투를 가진 기계 번역 모델보다 더 높은 성능을 기대할 수 있음을 보여준다.

5. 결론

이 연구는 ChatGPT를 이용한 한국어 캡션 생성에 대한 새로운 접근법을 소개하고, 이를 통해 큰 규모의 한국어 멀티모달 데이터셋을 더 효율적으로 만드는 방법을 제안한다. 지금까지 한국어 시각-언어 멀티모달 데이터셋의 부족함과 퀄리티 높은 데이터를 얻기 어렵다는 문제가 있었는데, 이번 연구를 통해 그 문제를 해결하고자 한다. 우리는 영어로 작성된 시각-언어 데이터를 ChatGPT를 사용하여 자연스럽고 고품질의 한국어 캡션으로 변환하는 새로운 방법을 모색한다.

실험 결과, ChatGPT 기반의 방식이 다른 번역 도구들에 비해 뛰어난 성과를 보여주고, 생성된 캡션은 이미지의 의미를 더 유사하게 전달한다는 것을 증명한다. 우리는 품질 좋은 한국어 데이터셋을 빠르게 구축할 수 있는 방안을 제시하며, 이 방식으로 구축된 데이터 셋을 통해 딥러닝 기반 한국어 Vision-Language Pre training 모델의 성능을 높일 수 있다. 따라서, 이 연구는 한국어 멀티모달 데이터셋 구축과 관련 분야에서 새로운 가능성을 보여줄 것이다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2022-00187238, 효율적 사전학습이 가능한 한국 어 대형 언어모델 사전학습 기술 개발).

참고문헌

[1] Chen, FL., Zhang, DZ., Han, ML. et al. VLP: A Survey on Vision-language Pre-training https://doi.org/10.1007/s11633-022-1369-5

[2] Reimers, Nils and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Conference on Empirical Methods in Natural Language Processing (2019).

[3] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[4] Hyunjoong Kim, "KoBERTscore", Github repository, https://github.com/lovit/KoBERTScore

[5] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[6] H. Agrawal et al., "nocaps: novel object captioning at scale," 2019 ICCV, 2019, doi: 10.1109/ICCV.2019.00904.

[7] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation. ACL2002

[8] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, ACL.