COSE474-2023F: Final Project Report
Testing the Korean Multimodal Abductive Reasoning Ability of KoCLIP

Jongbin Won

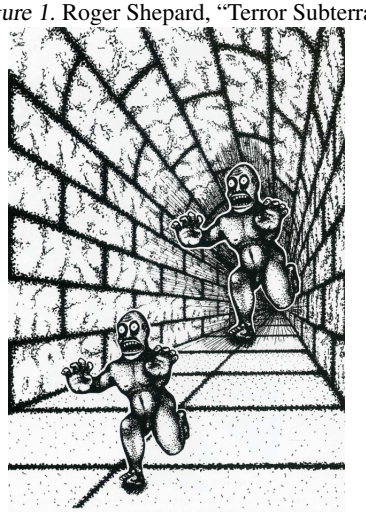
1. Introduction

1.1. Motivation

CLIP 모델은 시각과 언어 정보를 모두 활용해, 파인튜닝 과정 없이도 충분히 좋은 Zero-shot 성능을 보여주었다. 특히 CLIP 모델은 비행기, 인공위성과 같은 전문적인 영역보다는 일상에서 보기 쉬운 이미지들에 대하여 더 좋은 성능을 보여주며, 일상적인 상황들에 대한 이해도가 높다고 여겨져 왔다.(Radford et al., 2021)

그러나 인공지능 모델이 정말로 일상적인 상황들을 잘 이해한다고 주장하기 위해서는 단순히 해당 상황 속 객체와 텍스트간의 유사성을 파악하는 것 뿐만 아니라, 주어진 정보들을 활용해 해당 상황에 대한 판단 혹은 설명을 귀추 추론(Abductive Reasoning)해낼 수 있어야한다.(Zhao et al., 2023) 일례로 아래 사진을 보고, 인간은 단번에 앞의 사람이 뒤의 거인에게 쫓긴다는 상황을 추론해낼 수 있지만, 그러한 능력이 없는 인공지능은 단순히 해당 사진에 거인이 두 마리 있으며, 터널 속에 있다는 시각적 정보만을 파악할 수 있을 것이다.(Choi, 2022)

Figure 1. Roger Shepard, "Terror Subterra"



이처럼 시각-언어정보를 모두 활용하는 인공지능의 귀추추론 능력을 평가하는 연구들 중 영어에 기반한 연구들은 종종 있어왔으나, 한국어에 기반한 경우는 거의 찾아보기 어려웠다. 자연어 데이터가 인공지능 모델이 상식을 획득하는 핵심적인 단서라는 점에서, 인공지능 모델이 어떤 언어로 학습되었는지 여부는 해당 언어권의 사회, 문화를 이해할 수 있는지 여부를 결정하기에 매우 중요하다고 할 수 있다.(Choi, 2022) 따라서 본 프로젝트에서는 한국어로 재학습된 KoCLIP 모델을 활용해, 한국어와 시각정보에 기반한 인공지능 모델의 귀추추론 능력을 평가하고, 해당 능력을 향상시킬 수 있는 방법을 모색하고자 한다.

1.2. Problem Definition & Contributions

본 프로젝트는 구체적으로 다음과 같은 질문들에 대하여 답하고자 한다.

1) 귀추추론 능력 평가: KoCLIP 모델은 해당 상황이 담긴 사진과 문장 시퀀스를 활용해 귀추추론 과제를 수행할 수 있는가?

2) 귀추추론 능력 향상: 만약 가능하다면, KoCLIP 모델의 귀추추론 능력을 향상시킬 방법이 있는가?

이상의 연구를 통해 한국어와 시각 정보를 모두 받는 최신 모델로써 KoCLIP의 한국어 귀추추론 능력을 평가하고, 만약 해당 능력이 부족하다면 보완할 방법을 제시하는 것을 목표로 한다.

2. Methods

2.1. Datasets & Task

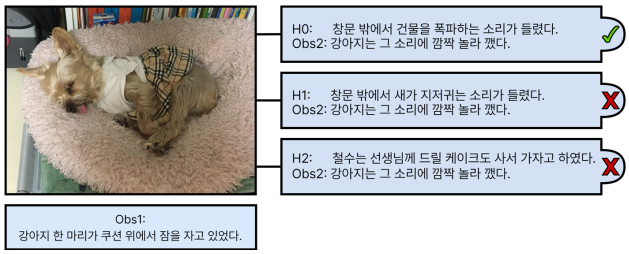
본 프로젝트는 국립국어원 주도로 구축된 한국어 귀추추론 데이터셋을 활용한다. 본 데이터셋은 총 12,284개의 데이터 샘플들로 구성되어 있다. 각각의 데이터 샘플은 하나의 이미지와 후술할 5개의 문장(Obs1,Obs2, Hyp0, Hyp1, Hyp2)으로 구성되어 있다. Obs1은 이미지에 대한 묘사이며, Obs2는 특정 상황이 발생한 뒤의 결과에 대한 묘사이다. 그리고 Hyp는 시간순으로 볼 때, Obs1과 Obs2 사이에 들어가 두 문장간의 연결을 자연스럽게 만드는 역할을 한다.

Hyp에는 총 3가지 선택지(Plausible, Implausible, Random)가 있으며, 이때 Plausible은 인간이 생각하기에 Obs2라는 결과를 야기하는 가장 그럴듯한 선택지이며, Implausible은 비록 논리적으로 불가능하지는 않지만 Plausible 보다는 일어날 가능성이 낮은, 덜 그럴듯한 선택지라고 할 수 있다. 한편, Random은 주어진 Obs1, Obs2와 무관한 선택지이다. 따라서 인공지능 모델이 인간만큼 귀추추론을 잘 할 수 있다면, Plausible, Implausible, Random순으로 많이

선택해야 할 것이다. 이때 선택지를 모델에게 입력값으로 넣을 때에는 Hyp0, Hyp1, Hyp2로 무작위로 할당하여 입력하였다.

이상의 데이터셋을 활용해 인공지능 모델의 한국어 귀추추론 능력을 평가하고자, 다음과 같이 태스크를 구성하였다.

Figure 2. Multiple Choice Task for Visual Abudctive Reasoning



인공지능 모델은 이미지와 3가지 텍스트 시퀀스들(Hyp0, Hyp1, Hyp2)을 입력으로 받는다. 각각의 텍스트 시퀀스는 Obs1과 Obs2의 상황을 HypN이 연결하고 설명해주는 Obs1+HypN+Obs2의 형태로 구성되어 있다. 그리고 주어진 언어와 시각 정보들을 종합할 때, 3가지 선택지 중 가장 그럴듯한 선택지를 선택해야 한다. 구체적인 슈도코드 알고리즘은 다음과 같다.

Algorithm 1 Inference Algoritms

$Image, Text\ Sequences = Data\ Sample$
$F_{image} = ImageEncoder(Image)$
$F_{text} = TextEncoder(Text)$
$Inputs = F_{image} \cdot F_{text} * \exp(temperature)$
$Output = Model(Inputs)$
$Predicted\ Label = argmax(Output)$

2.2. Environments

실험 환경은 다음과 같다. 라이브러리 및 기타 환경에 대한 더 상세한 정보는 Docker Hub에 공유해 둔 Docker Image를 불러와 확인할 수 있다.

- OS: Ubuntu 20.04. LTS
- CPU: intel i7-9700k 3.60GHz
- GPU: RTX 2080Ti * 2 (CUDA 11.7.99)
- Version details: Python 3.8.10, Pytorch 2.0.1

해당 환경에서 KoBERT와 KLUE-RoBERTa 모델의 경우, 8의 배치 사이즈로 각각 5에포크씩 파인튜닝시켰다. 파인튜닝 성능을 확인하고자 학습용 데이터셋의 20%를 검증용 데이터셋으로 할당하고, 500번째 스텝마다 검증용 데이터셋에 대한 정확도를 평가하였다. 그리고 검증용 데이터셋에 대한 정확도가 가장 높게 나온 모델 체크포인트를 활용해 최종 성능을 평가하였다. 최종 성능 평가에는 학습 및 검증용 데이터셋으로 사용되지 않은 1,216개의 테스트 데이터 샘플들을 활용했다.

한편, KoCLIP 모델의 경우 별도의 파인튜닝은 진행하지 않았으며, 텍스트 모델들과 같은 테스트 데이터 샘플들을 활용해 최종 성능을 평가하였다.

3. Results

이상의 과제에 대하여 KoCLIP모델을 활용해 Zero-shot으로 성능을 평가한 결과는 다음과 같다.

Table 1. Accuracy Comparing by Models & Fine-tuning

Model		Accuracy	
		Zero-shot	Finetuned
Text + Image	**KoCLIP**	**0.31907**	
Text-only	KoBERT	0.19408	0.67352
	KLUE-RoBERTa	0.20066	0.92434

파인 튜닝없이 테스트셋에 대한 Zero-shot 성능을 평가한 결과, KoCLIP은 약 31.907%의 정확도를 보여줬다. 이는 각각 약 19.408%와 20.066%의 Zero-shot 성능을 보여준 KoBERT 및 KLUE-RoBERTa 모델과 비교할 때, KoCLIP 모델에 비하여 10%가량 높은 Zero-shot 성능을 보여주었음을 확인할 수 있다. 단, KoCLIP 모델은 언어 정보 외에도 이미지를 추가적인 인풋으로 받는다는 점에서, 언어 정보만을 활용하는 KoBERT 및 KLUE-RoBERTa 모델과 직접적인 비교는 어렵다고 할 수 있다.

한편, 한국어로 시각기반 귀추추론을 진행하는 연구는 거의 없기 때문에 최신 모델과 직접적으로 정확도를 비교하기는 쉽지 않다. 그러나 영어를 기반으로 시각기반 귀추추론 문장 생성을 진행한 연구(Liang et al., 2022)에 따르면, 해당 논문이 제시한 모델(REASONER)은 BLEU@4 score 와 BERT score에서 각각 3.44점, 30.64점을 기록했다. 이는 같은 과제를 사람이 수행했을 때보다 훨씬 낮은 점수로, 아직 인공지능 모델의 귀추추론 능력이 사람보다 많이 부족한 것을 확인할 수 있다.

4. Experiments

4.1. Preprocessing Image & Prompt Engineering

CLIP 논문(Radford et al., 2021)에서 프롬프트 엔지니어링을 통해 5%p가량 정확도를 향상시켰던 점에서 착안하여, KoCLIP의 Zero-shot 성능을 향상시키고자 몇 가지 실험을 진행하였다.

먼저 이미지의 크기를 32*32사이즈로 조절하거나 표준정규분포로 정규화하는 등의 이미지를 전처리 한 뒤 귀추추론 정확도를 평가한 결과는 다음과 같다.

Table 2. Accuracy Comparing by Image Preprocessing

Preprocessing	Accuracy
No Preprocessing	**0.31907**
Normalization	0.29194
Normalization & Resize	0.27713

이미지 전처리와 관련해서는, 의외로 어떠한 전처리를 거치지 않은 경우의 정확도가 가장 높은 것으로 확인되었다. 이러한 현상은 KoCLIP을 사전학습 시키는 과정에서 정규화 등의 변형(transform)과정을 거치지 않은 이미지를 사용했거나, 추론 과정에서 이미 정규화하는 코드가 포함되어 있기 때문이라고 추측해볼 수 있다.

이미지 전처리를 통해서는 KoCLIP 모델의 추론 능력을 향상시키지 못한 한편, KoCLIP의 텍스트 인코더 부분을 담당하는 KLUE-RoBERTa 모델이 MLM(Masked Language Modeling)과 NSP(Next Sentence Prediction) 방식으로 사전학습되었다는 점(Park et al., 2021)에서 착안하여 다음과 같이 입력 텍스트 프롬프트를 수정하고 추론능력을 평가해보았다.

Table 3. Accuracy Comparing by Input Prompts

	Input Prompt	Accuracy
with Obs1	Obs1 + HypN + Obs2	0.31907
	Obs1 + [그 후] Obs2 + [그 이유는] HypN [기 때문이다]	0.32565
	Obs1 + [그리고] HypN + [그 결과] Obs2	0.31990
without Obs1	HypN + Obs2	0.36513
	[사진 속 장면 후에] + Obs2 + [그 이유는] HypN [기 때문이다]	0.35690
	[사진 속 장면 후에] + HypN + [그 결과] Obs2	**0.37006**

그 결과 Obs1을 사진으로 대체하고 적당한 접속어와 함께 Obs2-HypN 순으로 프롬프트를 배치해, 6%p가량 향상된 약 37.00%의 정확도를 확인할 수 있었다. 이러한 성능 향상은 텍스트 인코더가 사전학습된 방식과 유사하게 입력값을 넣어주면서도, 인풋 시퀀스의 길이를 줄였기 때문이라고 추측해볼 수 있다.

5. Conclusion

본 프로젝트는 한국어와 시각 정보를 모두 활용하는 인공지능 추론 모델에 관한 연구가 거의 없음을 인식하고, 한국어와 시각 정보를 모든 활용하는 인공지능 모델로써 KoCLIP의 귀추추론 능력을 파악하고 개선하는 것을 목표삼았다.

이를 위하여 한국어 귀추추론 데이터셋을 기반으로 다중택일 과제를 구성하였으며, KoCLIP 모델은 해당 과제에 대하여 약 31.90%의 성능을 확인할 수 있었다. 그리고 해당 모델의 귀추추론 성능을 향상시키고자, 이미지 전처리와 프롬프트 엔지니어링을 실험해보았다. 그 결과, 이미지 전처리는 성능을 향상시키지 못했으나, Obs1을 이미지로 대체하고 HypN과 Obs2사이에 적절한 접속어를 삽입하는 프롬프트 엔지니어링을 통하여 귀추추론 성능을 약 6%p 가량 향상시켜 최종적으로는 약 37.00%의 정확도를 확인할 수 있었다.

5.1. Discussions

KoCLIP은 많은 과제들에 대하여 훌륭한 Zero-shot 성능을 보여준 최신의 모델임에도 불구하고, 귀추추론 과제에 대해서는 최종적으로 약 37.00%의 정확도를 보여주었다. 이는 귀추추론 과제가 인공지능 모델에게 아직은 어려운 과제이며, 추가적인 연구가 필요하다는 것을 시사한다. 특히, 한국어와 이미지를 모두 활용하는 인공지능 모델이 극히 적은 만큼, 한국어 문화권의 문화/상식을 이해하는 귀추추론 모델에 대한 연구가 더욱 중요하다고 할 수 있다.

또한, 앞서 살펴보았듯 인공지능 모델을 추가로 학습시키지 않고, 추론 과정에서 단순히 입력 프롬프트를 조정하는 것만으로도 성능을 최대 6%가량 향상시킬 수 있었다. 따라서 인공지능 모델로부터 효율적인 정답을 이끌어낼 수 있는 프롬프트에 대한 연구 역시 추가적으로 진행되어야 할 것이다.

마지막으로 최근의 인공지능 모델들은 사전 학습단계에서 무수히 많은 데이터를 학습한 뒤, 파인튜닝없이 과제에 직접 사용되는 경우가 많다. 따라서 한국어 인공지능 모델이 별도의 파인튜닝 없이도 귀추추론 과제를 잘 수행할 수 있도록 한국어 귀추추론을 다루는 대규모 데이터셋을 구축할 필요가 있을 것이다.

5.2. Limitations & Further Researchs

본 연구의 한계점은 다음과 같다. 우선, 제시한 다중택일 과제가 인공지능의 귀추추론 능력을 충분히 잘 평가하는지에 대한 검증이 부족하다. 관련해서는 다른 데이터셋 및 과제를 통해 연관성을 확인하는 추가적인 작업이 필요하다.

다음으로, 시간상의 문제로 KoCLIP에 대한 파인튜닝 후 성능을 평가하지 못했다. 물론 CLIP류 모델의 경우, 파인튜닝을 통해 성능이 향상되지 않거나 되려 떨어진다는 연구도 있어왔다. 그러나 반대로 적절한 하이퍼파라미터 튜닝을 통해 CLIP류 모델의 파인튜닝 후 성능을 향상시킨 연구(Dong et al., 2022)도 있을 뿐더러, 본 과제의 경우 귀추추론이라는 특정한 태스크로 Domain-shift가 발생한 만큼 파인튜닝 후 성능 역시 추후 평가할 필요가 있다.

마지막은 학습 및 성능 평가를 위해 사용한 데이터 샘플의 수가 그리 많지 않다는 것이다. 총 12,284개의 데이터 샘플들을 검증용, 테스트용으로 따로 나누는 과정에서 훈련에 활용된 데이터 샘플들의 수가 줄어듦과 동시에, 최종 성능을 평가하기 위한 데이터셋 역시 1,216개로 충분히 확보하지 못했다. 추후 연구 과정에서는 추가적인 데이터셋 구축하는 등 데이터셋을 증강할 필요가 있을 것이다.

References

Choi, Y. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2):139–155, 05 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01906. URL https://doi.org/10.1162/daed_a_01906.

Dong, X., Bao, J., Zhang, T., Chen, D., Gu, S., Zhang, W., Yuan, L., Chen, D., Wen, F., and Yu, N. CLIP Itself is a Strong Fine-tuner: Achieving 85.7% and 88.0% Top-1 Accuracy with ViT-B and ViT-L on ImageNet, December 2022.

Liang, C., Wang, W., Zhou, T., and Yang, Y. Visual abductive reasoning. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J.-W., and Cho, K. Klue: Korean language understanding evaluation, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Zhao, W., Chiu, J. T., Hwang, J. D., Brahman, F., Hessel, J., Choudhury, S., Choi, Y., Li, X. L., and Suhr, A. Uncommonsense reasoning: Abductive reasoning about uncommon situations, 2023.