생성형 AI 의 교육용 컨텐츠 활용을 위한 연구

이승렬 [1], 오태훈 [2]
[1] 서경대학교 산업경영시스템공학과 학부생, [2] 한남대학교 글로벌비즈니스학과 학부생

icebear4825@skuniv.ac.kr, 20172207@gm.hannam.ac.kr

Research on the use of educational content in generative AI

Lee-Seung Ryul[1], Oh-Tae hoon[2]
[1]Dept. of Industrial Management & Systems Engineering, Seo-Keyong University
[2]Dept. of Global Business, Han-Nam University

요 약

본 논문에서는 LLM(Large Language Model) 모델의 fine-tuning 을 통한, 기초 수리 서술형 문항 풀이용 모델 및 Dall-E2 등 이미지 생성형 모델을 활용한 따른 영어 퀴즈풀이용 이미지 생성형 모델을 생성하여, 한국어 기반 LLM 자체 모델 학습 및 교육용 이미지 생성에 대한 방법을 고찰하였다.

1. 서론

생성형 인공지능(Generative Artificial Intelligence)은 인공지능 분야에서 중요한 발전을 나타내는 개념이다. 이는 인공지능 모델이 자율적으로 데이터를 생성하는 데 초점을 맞추고 있으며, 주로 텍스트, 이미지 또는 기타 멀티미디어 콘텐츠 형태로 이루어진다.

그 중에서도 언어 모델(Language Model)은 자연어처리 분야에서 핵심적인 역할을 하는 생성형 AI 의 중요한 구성요소이며, 생성형 AI 를 활용한 언어 모델과 교육의 결합은 학습자의 언어 습득, 문제 해결 능력 향상 등 다양한 교육 목표를 달성하는데 도움을 줄 것으로 예상되며, 이러한 결합은 학습자 중심의 교육 방법론을 더욱 강화하고, 학습자들이 보다 효과적으로 지식을 습득하고 응용할 수 있도록 지원한다.

본 논문에서는 언어 모델과 교육 분야 간의 긴밀한 관련성을 탐구하고, 생성형 AI 가 교육 분야에서 혁신적인 역할을 어떻게 수행하는지에 대해서 고찰한다.

2. 데이터셋

본 논문에서는 수학학습을 위한 LLM 모델 finetuning 과 영어 학습을 위한 이미지 생성형 모델, 위 두 가지 모두 초등학생을 대상으로 설정하였으며 이에 따라 학습가능한 모델을 선정하고 이에 최적화된 데이터를 생성하였다.

이를 위해 수학관련 문답 모음과 영어 이미지 생성을 위한 특정 키워드를 수집, 수학은 OpenAI 의 grade-school-math 의 train.json 파일을 활용하였으며, 총 7473 개의 문답 문항을 한국어로 번역하고, alpaca-lora dataset 에 맞추어 3 가지 속성(input, instruction, output)으로 나누어 전처리를 수행하였다. 이후 수학모델의 경우, 한국어 번역에 deep translator 를 활용하였다.

Text-to-image 모델의 경우, 영어 학습용 텍스트 데이터를 추리기 위하여, 교육부 지정 초등 영어단어 800 단어의 영단어를 활용하였으며, 이때 이미지 생성을 위한 text-prompt 를 생성할 시, 문법에 따른 구분이 필요하다고 판단하였고, 자연어 처리 toolkit 인 NLTK 를 활용하여, 800 단어 중 보통명사 312 단어를 이미지 생성을 위한 주요 텍스트로 지정하였다.

3. 모델 구현

Llama 2[2] ‘Large Language Model Meta AI’의 약어로 Meta 에서 공개한 대규모 언어모델이다.
Llama 의 경우는 2023 년 2 월에 공개되었으며, 이후 올해 7 월 Llama2 가 공개되었다. 이번 Llama2 는 매개변수 규모에 따라 세가지 모델(70 억개, 130 억개, 700 억개)로 공개되었으며 기존 보다 많은 2 조개의 토큰으로 학습되었으며, 대화형 모델인 Llama-2-Chat 의 경우 reinforcement learning with human feedback(RLHF)를 통하여 인간의 선호도에 부합하는 유용성과 안정성을 미세 조정한 하여 보다 인간 친화적인 모델이다.

Llama2 를 학습 모델로 넣은 이유는 일전에 언급하였던 Alpaca 등 수많은 파생형 모델을 생성하였던 것이 Llama 였고 이에 보다 뛰어난 성능의 오픈소스 모델로 Llama2 가 공개되어 이에 따라 한국어 및 영어 원본 데이터에 따른 Llama2 의 성능 실험을 진행

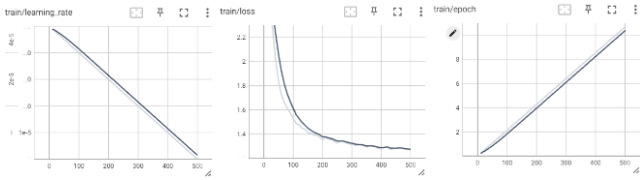
하였다.

또한 Polyglot-ko[1]는 한국어에 특화된 언어모델로, 1.2TB 의 대규모 한국어 데이터 셋을 사용하여 학습되었으며 Eleuthe AI 의 GPT-NeoX 를 기반으로 학습되었으며, Stability AI 의 HPC 클러스터에서 256 A100 을 활용하여 학습되었다 현재 Polyglot-ko 의 경우, 한국어 기반 모델 중 뛰어난 기량을 보이고 있으며, 지금 가지고 있는 데이터 셋을 활용하여 한국어 기반 모델을 학습 가능하다고 판단하였다.

Dall-E2[3]는 open ai 에서 개발한 image 와 텍스트를 연결해주는 neural network CLIP 에 기반한 Text-to-image 모델로, 텍스트로 CLIP image embedding 을 생성하고 이 embedding 을 condition 으로 받아 diffusion model 로 이미지를 생성하는 방식이다. 특히 open ai 의 open api 를 활용하여 서비스로 연동하기 편리하다는 점과 사실적인 이미지 생성에 적합하기에 해당 모델을 활용하여 이미지를 생성하고자 하였다.

4. 실험 결과

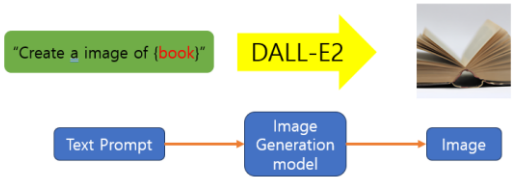
영문데이터와 한글데이터 두 데이터를 개별적으로 학습하였으며, Colab 을 활용하여, A100 GPU 한 대로 학습하였다. 영문 데이터의 경우, Lama2 7B 모델을 기반으로 학습하였으며, 학습 간 hugging face 의 노코드 기반 모델 학습용 도구인 Auto Train Advanced 를 이용하여 학습하였고, 한국어 데이터의 경우 polyglot-ko 13B 모델을 기반으로, 학습하였다. 학습한 가중치의 경우, (https://huggingface.co/re2panda)에서 확인할 수 있다.



<표 1> Polyglot12B 모델 기반 학습 결과

주요 모델인 Polyglot 의 경우, 500epoch 간, training loss 1.2661 으로, 유사한 구조의 데이터를 학습함으로써, dataset 에 fitting 되었다는 것을 알 수 있었고, 학습을 5e-05 로 학습할 시, 위와 같은 그래프를 보였으며 손실감소가 적절하게 이루어졌음을 알 수 있다.

이미지 생성형 모델의 OpenAI 의 Open API 를 활용하여 Dall-E2 를 활용하였으며, 텍스트 프롬프트를 생성하고자 하는 이미지 키워드를 제외하고 고정함으로써, 각 키워드 당 1024x1024 의 이미지 한 장을 base64 로 인코딩 한 후, 다시 이미지데이터로 디코딩하는 단계를 거쳐 생성하였다.



<그림 1> 생성형 이미지 생성 과정 예시

5. 제약

수학 모델을 학습함에 있어서 국내 수학 교재 내의 서술형 문항을 추출하여 학습데이터를 늘이는 방향 또한 생각하였으나. 저작권의 문제로 추가 데이터를 수집하지 못하고 학습을 진행하였다. 또한 Polyglot 이 한국어 기반 모델 중 학습 가능한 기반을 가졌기에 한국어 텍스트를 생성함에 있어서 자연스러운 문답을 내놓을 순 있었으나 수리적인 해설에서 부분적인 오류를 범하는 경우를 빈번하게 확인할 수 있었다. 이미지 생성형 모델의 경우, 특정 문법을 활용하여 이미지를 생성하였으나 인물을 묘사하는 단어 등 특정 단어에 대해서는 학습대상자가 단어를 추론할 수 없는 이미지를 생성하는 경우도 있었다.

6. 결론

본 논문은 두 갈래의 생성형 인공지능 모델을 활용한 교육 산업에서의 활용 방안을 제시하였다. 현재 데이터를 추가하여 14938 문항을 통한 학습 및 가중치 수정을 통하여, Benchmark 성능비교를 진행할 예정이며, 이를 통해 모델의 상용화 가능한 수준까지 높이고자 한다. 이미지 생성형 모델 또한, 이미지 모델 활용[4] 및 보다 정밀한 키워드 분류와 텍스트 프롬프트 세분화를 통하여, 보다 정밀한 교육용 자료를 생성하는 데에 기여할 수 있을 것으로 기대된다.

※ 본 논문은 과학기술정보통신부 정보통신창의-인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] Ko, Hyunwoong, et al. "A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models." arXiv preprint arXiv:2306.02254 (2023).
[2] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
[3] Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.
[4] Borji, Ali. "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2." arXiv preprint arXiv:2210.00586 (2022).