생성형 AI 기반 개인 맞춤형 동화책 제작 서비스 설계 및 구현

이소현, 원주연, 권용현, 김재호*

세종대학교

sohyun.sejong@gmail.com, juyeon.sejong@gmail.com,

yonghyun.sejong@gmail.com, *kimjh@sejong.ac.kr

Design and Implementation of Generative AI-based Personalized Storybook Creation Service

Sohyun Lee, Juyeon Weon, Yonghyun Kwon, Jaeho Kim*

Sejong Univ.

요 약

최근 생성형 AI 기술의 발전으로 생성형 AI로 생성된 텍스트, 이미지, 오디오 등의 디지털 콘텐츠는 다양한 분야에서 응용되며 여러 산업에 혁신적인 변화를 일으키고 있다. 본 논문에서는 생성형 AI를 이용하여 아동 교육을 위한 개인 맞춤형 이미지 기반 영어 동화책 서비스를 제안하고 구현한다. 아동이 그린 그림을 기반으로 시청각 요소를 결합한 동화책을 제작하여 아동의 상상력과 창의력 증진에 기여하고, 텍스트-스피치(TTS, Text-to-Speech) 기술을 활용하여 언어 교육에서의 긍정적인 효과를 기대한다.

Ⅰ. 서론

최근 ChatGPT, Gemini 등 대화형 생성형 AI가 급부상하며 생성형 AI가 많은 주목을 받고 있다. 생성형 AI는 대규모 데이터셋을 기반으로 훈련된 딥러닝 모델을 사용하여 새로운 데이터를 생성하는 인공지능의 한 분야이다. 거대 언어 모델(LLM, Large Language Model) 기반의 ChatGPT, Gemini뿐만 아니라, 텍스트-이미지 생성 모델, 텍스트-오디오 생성 모델 등 다양한 생성 모델은 전반적인 산업에 혁신적인 변화를 불러왔다.

생성형 AI로 생성된 텍스트, 이미지, 오디오 등의 디지털 콘텐츠는 다양한 분야에서 활용되고 있다. 교육 분야에서의 생성형 AI의 적용은 개인 맞춤형 학습, 지능형 학습 시스템 등 개인화된 학습을 가능하게 한다[1]. 특히 아동 교육에서의 생성형 AI와 디지털 스토리텔링 기술 적용은 아동 교육의 필요 요소와 관심사에 맞는 개인화된 학습 자료를 생성한다. 이는 학습자의 인지 및 사회적 발달을 향상하는 개인 맞춤화된 학습 경험을 지원하며, 창의성, 비판적 사고, 의사소통 등 사회적 역량 향상에 기여한다[2]. 그럼에도 불구하고 아동 교육을 위한 텍스트, 이미지, 오디오 등의 생성 모델 활용은 여전히 부족한 실정이다.

본 논문에서는 아동 교육에서의 생성형 AI 활용 방안으로, 개인 맞춤형 이미지 기반 영어 동화책 서비스를 설계하고 구현한다. 아동이 그린 이미지를 이미지 캡셔닝 기술을 통해 해당 이미지를 설명하는 텍스트를 생성한다. 이미지 캡셔닝은 이미지 정보를 파악하고 이를 적절한 텍스트 정보로 생성하는 기술이다. 텍스트 생성 모델을 이용하여 동화 내용을 구성한다. 동화 내용을 기반으로 텍스트-이미지 생성 모델, 텍스트-오디오 생성 모델을 이용하여 동화 삽화와 효과음을 생성한다. 텍스트-스피치(TTS, Text-to-Speech) 모델을 사용하여 동화 내용에 대한 음성 파일을 제공한다. 마지막으로 생성형 AI 모델을 통해 생성된 동화의 모든 콘텐츠를 결합하여 동화책을 제작한다. 아동의 그림을 바탕으로 구성된 시청각 자료를 통해 아동의 상상을 확장하여 상상력과 창의성 향상에 기여할 것이라 예상한다. 또한 TTS 기능을 활용한 언어 교육의 긍정적인 효과를 기대한다.

Ⅱ. 시스템 설계

1. 시스템 아키텍처

그림 1은 본 시스템의 구성도이다. 사용자가 이미지를 입력하면, 이미지 캡셔닝 기술을 통해 해당 이미지를 설명하는 텍스트를 생성한다. 이 텍스트를 기반으로 동화 내용을 구성하고, 각 장면에 맞는 프롬프트를 생성한다. 이 프롬프트를 활용하여 동화 삽화, 효과음, 제목을 생성하고, TTS 기능을 구현한다. 마지막으로 동화 콘텐츠를 통합하여 웹페이지를 통해 사용자에게 완성된 동화책을 제공한다.

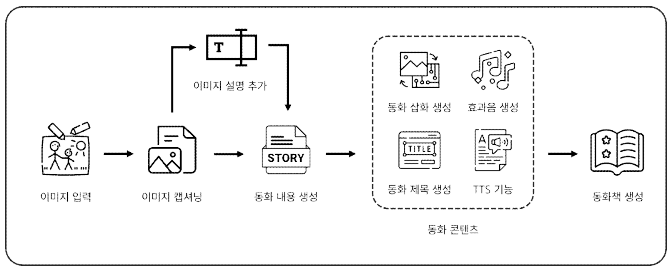

그림 1. 시스템 구성도

2. 세부 기능

본 시스템의 세부 기능은 다음과 같다.

2.1 이미지 캡셔닝

사용자가 입력한 이미지를 설명하는 텍스트를 생성하기 위해 비전 언어 모델인 BLIP을 사용한다. BLIP 모델은 새로운 비전-언어 사전 학습 프레임워크를 제안하며, 시각적 질문 답변, 이미지-텍스트 매칭, 이미지 캡셔닝 등에서 우수한 성능을 보인다[3].

2.2 동화 생성

동화 생성을 위해 텍스트 생성 모델인 GPT-3.5 Turbo를 사용한다. GPT-3.5 Turbo 모델은 추가 학습 데이터를 사용하여 GPT-3 모델을

파인튜닝한 모델이다. 사전 학습된 모델에 학습을 진행하여 다양한 자연어 처리 작업에 뛰어난 성능을 보이며, 이전 모델에 비해 자연스럽고 정확하며 일관된 문장을 생성한다[4].

2.3 동화 삽화 생성

각 장면의 텍스트 프롬프트를 이용하여 동화 삽화를 생성하기 위해 텍스트-이미지 모델인 Stable Diffusion을 사용한다. Stable Diffusion 모델은 잠재 확산 모델(LDM, Latent Diffusion Model)을 기반으로 한다. LDM은 노이즈를 효과적으로 제거하며 동시에 이미지의 품질을 유지하는 방식으로 학습 및 샘플링 효율성을 크게 향상시킨다. 따라서 Stable Diffusion은 이미지 합성 작업에서 뛰어난 성능을 보이며, 고품질의 사실적인 이미지를 효과적으로 생성한다[5].

2.4 동화 효과음 생성

각 장면의 텍스트 정보를 활용하여 동화의 효과음과 배경음악을 생성하기 위해 텍스트-오디오 생성 모델인 AudioLDM2를 사용한다. 이 모델은 LDM을 적용하여 입력된 텍스트 정보를 바탕으로 사실적인 음향 효과, 인간의 음성, 음악을 합성한다. AudioLDM2는 오디오 표현을 보편적 형태로 변환하는 자기 지도 사전 학습을 통해 강력한 오디오 생성 기반을 제공한다[6].

2.5 동화 TTS 기능

동화의 TTS 기능 구현을 위해 트랜스포머(Transformer) 기반의 텍스트-오디오 모델인 Bark를 사용한다. Bark 모델은 입력 텍스트 프롬프트가 음소를 사용하지 않고 오디오로 변환되기 때문에 발화 음성에 대한 사전 작업을 지원하며, 음악, 효과음, 비음성 소리의 처리가 가능하다. 또한 다국어 기능을 지원한다[7].

Ⅲ. 시스템 구현 결과

그림 2는 완성된 동화책 화면이다. 사용자가 입력한 이미지를 기반으로 동화책을 제작한다. 사용자가 그림을 입력하면 BLIP 모델을 사용하여 이미지 캡셔닝을 수행한다. 사용자는 이미지 캡셔닝 결과를 바탕으로 내용을 보완한다. “Create My own story book” 버튼을 클릭하여 동화책을 제작한다. 동화 내용, 동화 삽화, 효과음, TTS 기능이 순서대로 생성된다. 완성된 동화책 상단에는 GPT-3.5 Turbo 모델로 생성한 동화책 제목이 위치하며, 제목 하단에는 사용자가 입력한 이미지가 동화책 표지로 사용된다. 표지 하단에는 다섯 개의 장면으로 구성된 동화책 삽화와 내용이 위치한다. 동화책의 각 장면은 Stable Diffusion 모델로 생성한 두 개의 동화 삽화, GPT-3.5 Turbo 모델로 작성한 다섯 개의 영어 문장, AudioLDM2 모델으로 생성한 동화 장면의 효과음, Bark 모델을 이용한 TTS 기능으로 구성된다. 각 장면의 상단에는 해당하는 삽화가 배치되며, 하단에는 해당 장면의 동화 내용이 위치한다. TTS 기능으로 동화 내용을 순서대로 읽어주고, 효과음은 동화 내용에 맞추어 적절한 상황에 재생된다.

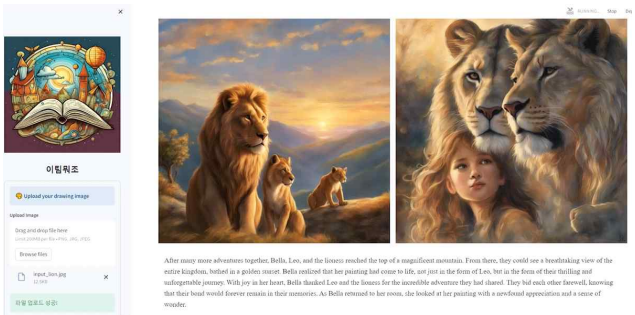

그림 2. 동화책 내용 웹페이지 화면
(https://www.youtube.com/watch?v=0gcZjo15lgw)

표 1. 기능별 사용 모델

기능	사용 모델	모델 주요 특징
이미지 캡셔닝	BLIP[3]	이미지-텍스트 생성
동화 내용 생성	GPT-3.5 Turbo[4]	텍스트-텍스트 생성
동화 삽화 생성	Stable Diffusion[5]	텍스트-이미지 생성
동화 효과음 생성	AudioLDM2[6]	텍스트-오디오 생성
TTS	Bark[7]	텍스트-오디오 변환

Ⅳ. 결론

본 논문에서는 생성형 AI를 활용하여 아동 교육을 위한 개인 맞춤형 이미지 기반 영어 동화책 서비스를 설계하고 구현하였다. 동화책 내용, 삽화, 효과음 등 동화책의 디지털 콘텐츠를 생성하기 위해 GPT-3.5 Turbo, Stable Diffusion, AudioLDM2 등 다양한 최신 생성형 AI 모델을 사용하였다. 본 시스템은 아동의 그림을 바탕으로 시청각 효과가 결합된 동화책을 제작한다는 점에서 아동의 상상력과 창의성 향상에 기여할 것으로 기대한다. 또한 다국어 지원이 가능한 TTS 기능을 활용하여 아동의 언어 교육의 긍정적인 효과를 기대한다. 향후 연구에서는 부모님, 친구 등 특정 인물의 목소리로 동화책을 읽어주는 TTS 기능을 개발하고자 한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업(IITP-2024-2021-0-01816)의 연구결과로 수행되었으며, 2024 년도 산업통상자원부 및 산업기술평가관리원 (KEIT) 연구비 지원에 의한 연구임 (RS-2022-00154678)

참 고 문 헌

[1] Baidoo-anu, D., & Owusu Ansah, L. "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning," Journal of AI, 7(1), 52-62, 2023.

[2] Baskara, FX Risang. "Fostering Culturally Grounded Learning: Generative Ai, Digital Storytelling, And Early Childhood Education," International Conference of Early Childhood Education in Multiperspectives. 2023.

[3] LI, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *In: Int. Conf. on machine learning. PMLR*, 2022, pp.12888-12900.

[4] Browm, Tom, et al. "Language models are few-shot learners," *Advances in neural information processing systems*, 33: 1877-1901, 2020.

[5] Rombach, Robin., et al. "High-resolution image synthesis with latent diffusion models," *In Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp.10684-10695.

[6] LIU, Haohe, et al. "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[7] Bark, Retrieved May. 03, 2024, from https://github.com/suno-ai/bark.