로봇 비전-언어 통합 기반 경험주의적 내러티브 생성 프레임워크

김유청, 김동민*
순천향대학교, *순천향대학교

yc424k@gmail.com, *dmk@sch.ac.kr

Empiricist Narrative Generation Framework via Robot Vision-Language Integration

Yu Cheong Kim, Dong Min Kim*
Soonchunhyang University, *Soonchunhyang University

요 약

본 연구에서는 이동 로봇의 실시간 시각 입력과 통신 인프라를 활용하여, 환경 경험을 바탕으로 문학적 서사를 생성하는 경험주의적 생성형 AI 시스템을 제안한다. 제안 시스템은 로봇이 촬영한 이미지를 고성능 이미지 캡셔닝 모듈로 요약하고, 이를 기반으로 장르별 파인튜닝된 거대 언어 모델이 서로 다른 문체의 내러티브를 동시 생성한다. 모더니즘 소설 스타일과 과학 소설 스타일로 분리된 두 개의 텍스트 생성기는 동일한 캡션을 받아 각기 독창적인 서사체를 출력하며, 시스템의 유연성과 확장성을 보장한다. 실험에서는 약 1,000 장의 로봇 시각 데이터와 3 개 코퍼스(모더니즘, 여행기, 과학 소설)를 사용하여 파이프라인을 구성하고, GPT-4o 기반 자동 평가를 통해 문장 품질, 어휘 다양도, 세계관 구축, 감정 표현, 응집성, 몰입도 등 6 가지 지표에서 성능을 분석하였다. 그 결과, 제안 모델은 기존 LSTM 기반 접근법 대비 모든 평가 항목에서 평균 15% 이상 우수한 성능을 보였으며, 문체 제어와 실시간 경험 반영의 가능성을 입증하였다. 본 연구는 로봇-AI 협력 환경에서 경험 기반 디지털 스토리텔링의 새로운 설계 방향을 제시하며, 향후 다중 감각 입력 및 대화형 내러티브로의 확장을 위한 기반을 제공한다.

Ⅰ. 서 론

현대 생성형 AI 연구는 주로 정적 데이터셋에 의존하여 텍스트를 생성하지만, 실제 세계와의 능동적 상호작용을 통해 생성된 데이터 기반의 서사 창작 가능성은 아직 충분히 탐구되지 않았다. Ross Goodwin(2018)의 『1 the Road』[1]는 이동형 신경망이 자동차 여행 경험을 문학적 텍스트로 전환한 대표적 사례이나, 시스템 구성과 문체 제어 측면에서 확장 여지가 있다. 이에 본 연구는 이동 로봇의 카메라와 무선 통신을 통합하여, 실시간으로 촬영된 이미지를 중앙 서버로 전송하고, 이미지 캡셔닝과 LLM 파인튜닝을 결합한 경험주의적 내러티브 생성 프레임워크를 제안한다.

Ⅱ. 본론

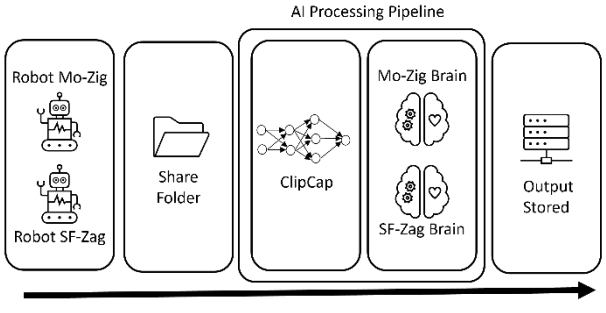

그림 1. 로봇 비전-언어 통합 기반 생성 프레임워크

2.1 전체 파이프라인 개요

그림 1 은 제안 시스템의 전체 동작 흐름을 보여준다. 로봇은 Mo-Zig(모더니즘 소설 스타일) 또는 SF-Zag(과학 소설 스타일) 역할을 할당받아 주변 환경을 이동하며 이미지 데이터를 수집한다. 수집된 이미지는 ZigBee 및 5G 네트워크를 통해 중앙 서버의 공유 폴더로 전송되며, 서버 내 리스너가 신규 파일 생성을 감지하면 AI 처리 모듈을 호출한다.

2.2 이미지 캡셔닝 모듈

본 연구에서는 ClipCap[2]을 기반으로 하여, 이미지-텍스트 임베딩 융합 능력을 활용한 캡션 생성 모델을 사용하였다. ClipCap 은 CLIP[3]의 사전 학습된 비전-언어 임베딩을 Prefix 로 활용하여, 이미지의 주요 객체와 배경 정보를 자연어 문장으로 요약한다. 데이터셋은 약 1,000 장의 로봇 촬영 이미지로 구성되었으며, 도로, 건물, 자연물 등을 포함한다.

2.3 문체별 LLM 생성기

캡션 입력을 받아 서사를 생성하는 두 개의 LLM 은 각각 GPT-2 기반으로 구현되었으며, DeepSpeed[4]를 활용한 분산 파인튜닝으로 대규모 코퍼스 학습을 효율화하였다. 코퍼스 구성은 다음과 같다:

- 모더니즘 소설 코퍼스 A: 212 편
- 여행기 소설 코퍼스 B: 100 편
- 과학 소설 코퍼스 C: 200 편

Mo-Zig 모델은 A+B, SF-Zag 모델은 C+B 조합으로 학습하여, 여행기 코퍼스를 문체적 브리지로 활용하였다. 파인튜닝 시 학습률, 배치 크기, 에포크 수

등 하이퍼파라미터를 그리드 서치 기법으로 최적화하였다.

Ⅲ. 실험 및 평가

촬영된 이미지를 기반으로 파인튜닝된 모델을 사용하여 서사를 생성하는 실험을 진행하고, 생성된 텍스트의 품질을 다각도로 평가하기 위해 GPT-4o 모델을 활용한 정량적 평가를 수행했다. 비교를 위해 LSTM 기반 문장생성기를 동일한 데이터셋으로 학습시켰다. 평가는 6 가지 주요 기준, 문장 품질, 캐릭터 개연성, 세계관 구축, 감정 표현, 응집성, 몰입도에 대해 각 항목별로 0 점에서 10 점 사이의 점수를 부여하는 방식으로 진행되었다, 평가 결과는 표 1 에 요약되어 있다.

표 1:GPT-4o 를 활용한 생성 텍스트 평가 결과

Model Type	Writing Quality	Character Plausibility	World Building	Emotional Expression	Coherence	Engagement
Modern LSTM	2.5	1.5	2.0	1.8	1.2	2.0
SF LSTM	2.5	1.0	2.0	3.0	1.5	2.0
Modern LLM	7.0	6.5	5.5	6.0	6.5	6.0
SF LLM	6.0	5.5	6.0	7.5	4.0	7.0
General Novel	8.5	7.8	9.2	9.0	7.5	8.3

분석 결과, LLM 기반 모델(Modern LLM, SF LLM)이 LSTM 기반 모델(Modern LSTM, SF LSTM)에 비해 모든 평가 항목에서 높은 점수를 받았다. 이는 LLM 이 LSTM 에 비해 문맥 이해, 문장 생성 능력, 그리고 복잡한 서사 구조 형성에서 더 뛰어난 성능을 보임을 시사한다.

또한, 일반적인 소설과 비교했을 때, 파인튜닝된 LLM 모델들은 여전히 일부 항목에서 개선의 여지를 보였으나, 특정 장르의 스타일을 반영하면서도 일정 수준 이상의 서사적 품질을 달성할 수 있는 가능성을 보여주었다. SF LLM 의 경우 '감정 표현'과 '몰입도'에서 Modern LLM 보다 높은 점수를 받았는데, 이는 SF 장르의 특성이 감정적 고조와 독자의 몰입을 유도하는 데 더 효과적으로 작용했음을 시사할 수 있다. 반면, '응집성' 항목에서는 SF LLM 이 상대적으로 낮은 점수를 받아, 복잡한 세계관이나 사건 전개에 따른 논리적 흐름 유지에 어려움이 있었을 가능성을 나타낸다. 이러한 결과는 특정 코퍼스를 통해 파인튜닝된 LLM 이 해당 코퍼스의 문체적 특성을 학습하고 발현할 수 있음을 보여주며, 동시에 파인튜닝 과정과 원본 LLM 의 내재적 특성이 상호작용하여 최종 생성물의 스타일에 영향을 미침을 시사한다. 코퍼스 설계를 통해 LLM 의 생성 스타일을 유의미하게 제어할 수 있는 가능성을 확인할 수 있다.

Ⅳ. 결론

본 논문에서는 이동 로봇으로부터 수집되는 동적인 시각 감각 입력과 특정 문학 장르 코퍼스를 활용한 거대 언어 모델의 파인튜닝을 효과적으로 결합함으로써, 고유한 문체적 특성을 지닌 개별화된 글쓰기 주체, 즉, '경험주의적 내러티브 생성 프레임워크'를 구현할 수 있는 가능성을 성공적으로 제시하였다. 그림 1 에 제시된 통합 파이프라인과 같은 시스템을 통해, 기계는 단순히 사전에 학습된 정보를 재생산하는 것을 넘어, 현실 세계를 실시간으로 '경험'하고 그 경험을 바탕으로 서로 다른 스타일과 관점을 담은 독창적인 서사를 생성할 수 있음을 실험적으로 입증하였다.

향후 연구에서는 다음과 같은 방향으로 본 연구의 접근 방식을 더욱 심화하고 확장할 필요가 있다. 첫째, 현재 시각정보에 국한된 로봇의 감각 입력을 청각, 촉각 등 다양한 멀티모달 정보로 확장하여 보다 풍부하고 입체적으로 경험 기반의 서사 생성을 가능하게 하는 연구가 필요하다. 둘째, 생성된 서사의 문학적 가치, 창의성, 그리고 독자에게 미치는 정서적 영향 등을 보다 심층적으로 평가할 수 있는 정교한 평가 지표 및 방법론 개발이 병행되어야 한다. 이러한 후속 연구들을 통해, 본 연구에서 제시한 경험주의적 내러티브 생성 시스템은 인공지능과 인간의 창의적 협력 관계를 새롭게 정립하고, 더욱 풍요로운 디지털 스토리텔링 시대를 열어가는 데 기여할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

이 논문은 2024 년 대한민국 교육부와 한국연구재단의 인문사회분야 일반공동연구지원사업(융복합연구)의 지원을 받아 수행된 연구임(NRF-2022S1A5A2A03052880). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학·석사연계 ICT 핵심인재양성사업의 연구결과로 수행되었음. (IITP-2025-RS-2024-00436500)

참 고 문 헌

[1] R. Goodwin, 1 the Road, Jean Boîte Éditions, 2018.

[2] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," arXiv preprint arXiv:2111.09734, 2021.

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. 38th Int. Conf. Mach. Learn. (ICML), 2021.

[4] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "ZeRO: Memory Optimization Toward Training A Trillion Parameter Models," in Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis (SC), 2020.

[5] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998- 6008.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1877- 1901.