프롬프트 기반 퓨샷 러닝을 통한 한국어 대화형 텍스트 기반 이미지 생성

이은찬°, 안상태*
경북대학교 전자전기공학부, 경북대학교 전자공학부
*stahn@knu.ac.kr

Image Generation from Korean Dialogue Text via Prompt-based Few-shot Learning

Eunchan Lee°, Sangtae Ahn*
School of Electronic and Electrical Engineering, School of Electronics Engineering

요 약

본 논문에서는 사용자가 대화 텍스트 방식의 입력을 주었을 때 이를 키워드 중심으로 변환하여 이미지를 생성해내는 방식을 제안한다. 대화 텍스트란 채팅 등에서 주로 사용하는 형식의 구어체를 말하며 이러한 텍스트 형식은 텍스트 기반 이미지 생성 모델이 적절한 아웃풋 이미지를 생성하기 어렵게 만든다. 이를 해결하기 위해 대화 텍스트를 키워드 중심 텍스트로 바꾸어 텍스트 기반 이미지 생성 모델의 입력으로 변환하는 과정이 이미지 생성의 질을 높이는 좋은 방안이 될 수 있는데 이러한 태스크에 적합한 학습 데이터는 충분하지 않다. 본 논문에서는 이러한 문제를 다루기 위한 하나의 방안으로 사전학습된 초대형 언어모델인 KoGPT 모델을 활용하며, 퓨샷 러닝을 통해 적은 양의 직접 제작한 데이터만을 학습시켜 대화 텍스트 기반의 이미지 생성을 구현하는 방법을 제안한다.

주제어: 텍스트 기반 이미지 생성 모델, 멀티 모달, 한국어 기반 초대형 언어 모델, KoGPT, 프롬프트 기반의 퓨샷 러닝

1. 서론

멀티 모달 모델링은 자연어 처리와 컴퓨터 비전 학계 양방향에서 최근 가장 주목받는 분야 중에 하나로 여겨진다. 대표적인 연구 분야로는 기존부터 높은 성능을 거두며 널리 응용되어 왔던 이미지-투-텍스트 태스크인 이미지 캡셔닝 (Image Captioning), 시각 질의응답(Visual Question Answer, VQA), 광학 문자 인식(Optical Character Recognition, OCR)를 들 수 있으며, 최근에는 텍스트-이미지 동시 이해와 이미지 생성을 결합하여 응용한 텍스트-투-이미지의 태스크인 텍스트 기반 이미지 생성 모델이 최근 들어 안정적이고 놀라운 성능을 보이며 화두가 되고 있다. 이러한 텍스트 기반 이미지 생성 모델링에는 대표적으로 OpenAI의 DALL.E 2와 Google의 Imagen을 들 수 있다.[1, 2]

이러한 텍스트 기반 이미지 생성 모델링은 모델 관점에서는 텍스트 입력을 이해하여 이를 문맥을 포함하는 은닉 벡터로 보내주어 이미지를 생성하게하는 인코더-디코더 모델을 활용하는 태스크로 볼 수 있다. DALL.E 2와 같은 대부분의 모델은 프롬프트라고 부르는 구체적인 명령문 형식의 텍스트를 입력으로 받아 생성하도록 학습된다.

그러나 AI를 전혀 알지 못하는 사용자가 텍스트 기반 이미지 생성 모델의 서비스를 이용하기 위해 접근할 때는 구체적인 명령문 형식의 프롬프트 형태의 텍스트 입력보다는 조금 더 날것의 구어체 형식을 지니는 대화형 텍스트 입력이 더욱 편리할 수 있다. 따라서 모델이 대화형 텍스트를 이해하여 이미지를 생성할 수 있도록 하는 기술의 수요가 충분히 있을 수 있는데 모델은 학습과정에서 대화형 텍스트를 이해하여 이미지를 생성하는 데에는 최적화 되지 않았기 때문에 모델의 성능과 별개로 대화형 텍스트의 문맥을 제대로 담지 못하는 좋지 못한 이미지를 생성해낸다는 문제를 지니게 된다.

그러므로 본 논문에서는 사용자의 구어체 대화형 텍스트를 입력으로 주었을 때 이를 자동으로 텍스트 기반 이미지 생성 모델의 입력에 적합한 키워드 기반의 프롬프트로 변환하여보다 문맥을 잘 담은 이미지를 생성하는 기법을 제시하고자 한다.

대화형 텍스트를 키워트 텍스트로 변환하는 과정은 텍스트 스타일 변환, 텍스트 요약과 유사한 면이 있으나, 학습을 위한 데이터셋이 충분하지 못하다는 단점이 있기 때문에 본 논문에서는 GPT-3 [3]의 연구에서 제시된 초대형 언어 모델을 적은 양의 데이터로 효과적으로 학습할 수 있는 프롬프트 기반의 퓨샷 러닝 방식인 인 컨텍스트 러닝을 한국어 초대형 언어 모델인 KoGPT[4]에 적용하여 학습하였다.

2. 관련 연구

2.1 GPT-3와 초대형 언어 모델

사전학습된 초대형 언어 모델 [3, 4, 5, 6, 7]은 언어 모델에 엄청 많은 양의 텍스트 말뭉치를 사전학습하여, 파인 튜닝 시에 더욱 높은 성능을 거둘 수 있는 장점이 있다. 사전학습 시 데이터의 양과 언어 모델의 모델 파라미터 양에 비례하여 다운스트림 태스크의 성능이 상승하는 경향이 있는데, GPT-3 [3]에서는 이를 극대화하기 위하여, 사전 학습 데이터셋의 양을 3,000억 개 까지 늘려 학습하여 1750억 개(175B)의 모델 파라미터의 모델과 그 외의 모델 사이즈들을 제시하였다.

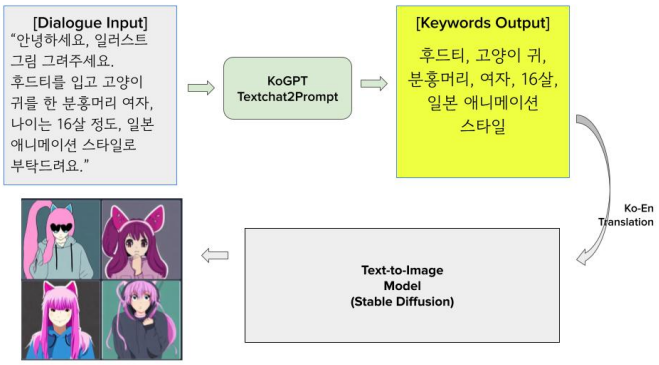

그림 1. Textchat2Prompt 기반의 대화형 텍스트 기반 이미지 생성 전체 과정의 개요

그리고 이렇게 고도로 사전학습된 모델이 적절한 양의 데이터셋을 필수로 요구하는 파인 튜닝이 아닌 적은 양의 데이터셋을 통해 학습할 수 있는 퓨샷 러닝을 통해서도 파인 튜닝에 뒤지지 않는 높은 성능을 보일 수 있음을 보여주었다.

관련 연구들에서는 통상적으로 10억 개(1B) 이상의 모델 파라미터를 가진 사전 학습된 초대형 언어 모델의 경우에 퓨샷 러닝을 통해 적절한 수준의 태스크 성능을 거둘 수 있다고 제시하였다. [3, 8]

한국어 기반 사전학습 초대형 언어 모델로는 Kakaobrain의 KoGPT [4]와 NAVER CLOVA의 HyperCLOVA [9]를 대표적으로 들 수 있다. KoGPT는 약 61억 (6.16B) 개의 모델 파라미터를 가지며, HyperCLOVA는 69억 개와 820억 개를 각각 모델 파라미터로 가지는 두 가지 버전이 공개되어있다. 그러나 KoGPT 모델은 완전히 오픈소스로 공개되어있다는 점에서 사용 시 여러가지 제약이 있는 HyperCLOVA에 비해 장점이 있다.

2.2 텍스트 기반 이미지 생성 모델

DALL.E [10]는 GAN 기반의 이미지 생성 태스크에 텍스트를 이해하는 방식을 결합하여, 제로 샷 텍스트 기반 이미지 생성이 가능함을 보여주였다. 이는 이후 다양한 텍스트 기반 이미지 생성 모델들이 연구될 수 있는 초석을 마련하였으며, DALL.E 2 는 Diffusion이라는 이미지 생성 방식을 적용하여 보다 현실적이면서 정량적, 정성적으로 더 높은 수준의 텍스트 기반 이미지 생성을 가능하게 했다.

최근에는 Google의 Imagen 또한 Diffusion 기반의 모델 기반의 텍스트 기반 이미지 생성 모델을 제시하여 화두가 되었으며, latent space 기반의 방식을 통해 Diffusion의 성능을 개선시킨 Latent Diffusion 모델 [11]기반의 모델인 Stable Diffusion [11] 모델이 높은 생성 수준의 텍스트 기반 이미지 생성 모델로 주목받고 있다. DALL.E 2 모델과 비교하였을때 Stable Diffusion 모델은 비슷한 성능에 완전히 오픈소스로 공개되어있다는 점에서 사용 시 결제가 필요한 DALL.E 2에 비해 장점이 있다.

3. Textchat2Prompt 기반의 대화형 텍스트 기반 이미지 생성

3.1 제안 방식

본 논문에서는 그림 1과 같이 언어 모델을 활용하여 대화 형식의 텍스트 입력을 키워드 형식의 텍스트로 변환하고, 이를 텍스트 기반 이미지 생성 모델의 입력 텍스트로 주어 이미지를 생성해내는 방식을 통해 대화형 텍스트 기반의 이미지 생성을 구현하였다. 최종 변환된 한국어 입력 텍스트를 기계 번역을 통해 영문으로 번역하여 텍스트 기반 이미지 생성의 입력으로 사용하였다.

3.2 Textchat2Prompt: 대화형 텍스트를 키워드 형식 텍스트로 변환

본 논문에서는 인-컨텍스트 러닝(In-context Learning)이라고도 불리는 GPT-3 [3] 논문에서 제시된 사전학습된 초대형 언어 모델을 프롬프트 기반의 Few-shot Learning을 통해 학습시키는 방법을 적용하여 대화형 텍스트를 키워드 형식의 텍스트로 변환하는 기능을 구현하였다.

이에 사용한 초대형 언어 모델로는 KoGPT를 사용하였으며, 이는 대용량의 한국어 말뭉치로 사전학습 되었다. KoGPT의 모델 파라미터는 약 60억 개를 가지므로 GPT-3의 6B 버전과 GPT-J와 유사한 모델 파라미터를 가지는 특징이 있으며, 모델 파라미터의 수 관점에서 프롬프트 기반의 퓨샷 러닝을 통해 충분한 성능을 내기에 적합하다고 볼 수 있다.

퓨샷 러닝을 위한 데이터셋 키워드를 추출해내는 태스크를 학습하기 위하여 대화형 텍스트-키워드 쌍의 데이터를 직접 생성하여 이를 기반으로 K=4의 퓨샷 러닝으로 초대형 언어 모델을 학습하였다. K개의 구축한 데이터 쌍 중 하나는 아래와 같으며 아래와 같은 데이터셋을 모델의 입력으로 주어 학습하는 일종의 메타 러닝 방식으로 모델이 입력을 통해 태스크를 학습할 수 있게 구성하였다.

입력: 웹툰 표지로 일러스트를 쓰고싶은데 그림 그려줘. 웹툰은 좀비 던전과 보물에 얽힌 비밀들에 관한 내용이고, 내가 필요한 일러스트는 좀비 던전을 지나는 칼을 든 금발의 5살 여자 아이야.

출력: 일러스트레이션, 웹툰, 좀비 던전, 칼을 든, 금발, 5살, 여자 아이

3.3 텍스트 기반 이미지 생성 모델

Stable Diffusion [11] 모델을 통해 이미지를 생성하였다. Stable Diffusion 모델은 state-of-the-art 급의 텍스트 기반 이미지

표 1. 대화형 텍스트 - 키워드 텍스트 변환 결과

	텍스트
예시1 입력	”그림을 그려주세요. 음 기타를 든 잘생기고 젊은 가수로 부탁하고, 배경은 대학교 축제로 해주세요!”
예시1 변환된 키워드	**대학교 축제, 기타를, 든, 잘생기고, 젊은, 가수, 배경**
예시2 입력	안녕하세요, 일러스트 그림 그려주세요. 후드티를 입고 고양이 귀를 한 분홍머리 여자, 나이는 16살 정도, 일본 애니메이션 스타일로 부탁드려요.
예시2 변환된 키워드	**후드티, 고양이 귀, 분홍머리, 여자, 16살, 일본 애니메이션 스타일**

생성 모델로 볼 수 있으며, 약 50억 개의 다국어 이미지-텍스트 쌍으로 학습되었다. 한국어 데이터를 완전히 이해하지 못하는 것은 아니지만 영어 기반 입력에 비해 이미지 출력 성능이 좋지 못했기 때문에 본 논문에서는 Stable Diffusion을 통해 이미지를 생성하기 위해 기계 번역 API를 이용한 번역 중간 과정을 거쳐 최종적으로 3.1을 통해 변환한 한국어 텍스트 키워드를 Stable Diffusion이 영문 입력으로 이해할 수 있도록 하였다. 번역에는 파파고 API [12]를 사용하였다.

4. 정성적 평가를 위한 대화형 텍스트 기반 이미지 생성 결과

본 논문에서는 일러스트 기반의 아웃풋을 제시하여 정성적으로 결과를 제시하고자 한다. 관련 벤치마크 태스크가 거의 없으며 한국어 기반의 벤치마크 역시 전무하다는 점이 있으며, 또한 언어 모델과 텍스트 기반 이미지 생성 모델 면에서 고성능의 SOTA급의 모델을 베이스라인으로 하였다는 점을 그 이유로 들 수 있다. 따라서, 본 논문에서는 정량적 평가를 제쳐두고 일러스트를 생성하기를 원하는 가상의 대화형 텍스트를 입력 데이터로 주었을 때 모델이 생성해내는 이미지가 일러스트에 적합한지를 명확히 보여주기 위하여 아웃풋 이미지를 정성적 평가 지표로 제시한다.

4.1 대화형 텍스트 - 키워트 텍스트 변환 결과

본 섹션에서는 KoGPT를 few-shot learning 하였을 때, 몇 가지 예시의 대화형 텍스트를 키워드 텍스트로 변환하는 Textchat2Prompt의 결과를 제시한다. 결과는 표 1과 같았다. 키워드 기반의 텍스트를 매우 잘 추출해냄을 볼 수 있었다.

4.2 이미지 생성 결과

본 섹션에서는 4.1에서 생성한 키워드 기반의 프롬프트 텍스트를 텍스트 기반 이미지 생성의 state-of-the-art 모델인 Stable Diffusion을 통해 생성한 이미지 결과를 제시하여 성능을 검증한다.

그림 2. 대화형 텍스트 Textchat2Prompt를 적용하지 않은 생성 이미지 **(위)**와 Textchat2Prompt를 적용하여 생성한 이미지 **(아래)** 결과

먼저, 그림 2는 표 1의 예시 1의 텍스트에 대한 결과를 보여주는 그림이다. 먼저 대화형 텍스트를 그대로 입력했을 때는 그림 2의 위 그림 처럼 정확한 요구사항을 파악하지 못하는 경향을 볼 수 있다. 반면에 예시 1의 대화형 텍스트에 포함되어 있는 ’그려주세요’, ’음’, ’부탁하고’, ’해주세요!’와 같은 노이즈

가 Textchat2Prompt를 거쳐서 제거됨에 따라 그림 2의 아래 그림이 대화형 텍스트에서 요구하는 조건을 더욱 명확하게 이해하고 아웃풋을 출력해냄을 볼 수 있었다.

또한, 그림 3은 표 1의 예시 2 텍스트를 Textchat2Prompt를 통해 키워드 형식으로 변환했을 때 나타나는 출력 그림을 정성적 평가로 제시한 것이다. '안녕하세요', '그림 그려주세요'와 같은 자연스럽지만 모델의 관점에서는 노이즈에 불과한 텍스트들이 모두 제거됨에 따라, 텍스트 기반 이미지 생성 모델이 키워드 중심으로 대화형 텍스트의 요구 조건에 완벽히 부합하는 명확한 이미지를 생성해내는 것을 정성적으로 확인해 볼 수 있었다.

5. 결론

본 논문에서는 이미지-텍스트 멀티 모달 기술의 핵심인 텍스트 기반 이미지 생성 모델의 서비스를 사용자가 이용하는 과정에서 발생할 수 있는 대화형 텍스트 입력을 잘 다루지 못하는 문제를 다루기 위해 초대형 언어 모델을 퓨샷 러닝을 통해 학습하는 방식 기반의 Textchat2Prompt를 구현하여 적절한 입력 방식인 키워드 형식의 입력으로 바꾸어 주어 사용자가 더욱 높은 수준의 텍스트 기반 이미지 생성이 가능하도록 하는 방식을 제시하였다.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A4A1023248).

그림 3. 표 1의 예시 2의 대화형 텍스트에 Textchat2Prompt를 적용했을 때 모델이 생성한 대화형 텍스트의 요구 조건에 모두 부합하는 이미지

참고문헌

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[4] I. Kim, G. Han, J. Ham, and W. Baek, "Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer," https://github.com/kakaobrain/kogpt, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] J. Lee, "Kcbert: Korean comments bert," *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pp. 437–440, 2020.

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

[8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[9] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo *et al.*, "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers," *arXiv preprint arXiv:2109.04650*, 2021.

[10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-

to-image generation," *International Conference on Machine Learning*, pp. 8821–8831, 2021.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[12] H.-G. Lee, J.-S. Kim, J.-H. Shin, J. Lee, Y.-X. Quan, and Y.-S. Jeong, "papago: A machine translation service with word sense disambiguation and currency conversion," *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 185–188, 2016.