StoryWeaver: Text-to-Image generative model for cartoonists

YBIGTA StoryWeaver Team

June 28, 2024

Team Members

- Kim YeRin 2020122034 ryn0715@naver.com Team Leader, LLM pipeline, frontend, presentation material, presentation
- Kim ChaeHyun 2021122028 kimchaehyun0315@yonsei.ac.kr LLM pipeline, frontend
- Nam SeHyun 2022149028 daniel5253@yonsei.ac.kr
- Seo GunHa 2019142143 gunhaseo@yonsei.ac.kr backend, frontend, database construction
- Yang InHye 2021195070 inhyeyang@yonsei.ac.kr dataset preparation, data processing, DreamBooth finetuning
- Lee SungHyun 2020114010 sheepswool@yonsei.ac.kr dataset preparation, data processing, DreamBooth finetuning
- Jung SuHyun 2020122039 pikachuisabird@yonsei.ac.kr

1 Problem Statement

The intensive cycle of workload is the major bottleneck for cartoonists [1] [2]. Cartoonists suffer both from mental and physical disease due to the high level of labor intensity [3]. There existed various efforts from Artificial Intelligence to relieve the burden of workers or replace certain processes of work [1]. For example, Naver offered a service that automatically paints the sketch based on Deep Learning technique in 2021, and Naver webtoon has released a service visualizing the user idea. However, Naver halted some related services due to copyrights [4] and the unsatisfactory result of the service complained by webtoon consumers. Such cases represent the major causes for bothering the utilization of AI-based services in the webtoon labor environment. We propose StoryWeaver, a text-to-image generative model that inputs a cartoon scenario from the writer and generates a storyboard consisted of a consistent sequence of images and their corresponding text. Storyboard, also known as conti, is a blueprint for drawing and organizing a cartoon.

2 Related work

2.1 Text to Image Generation

Txt2img is a multimodal task that generates images with the input text. In our task, we add one more task which follows conformity between generated images. We enforce the model to assume the sequence of texts guarantees high relation, and the generated images follow consistency. Text-to-image generation originated from image-to-image generation models, where Variational AutoEncoder (VAE) is widely used for the given task. Generative models such as GANs were first used as image-to-image models, which eventually expanded as text-to-image models.

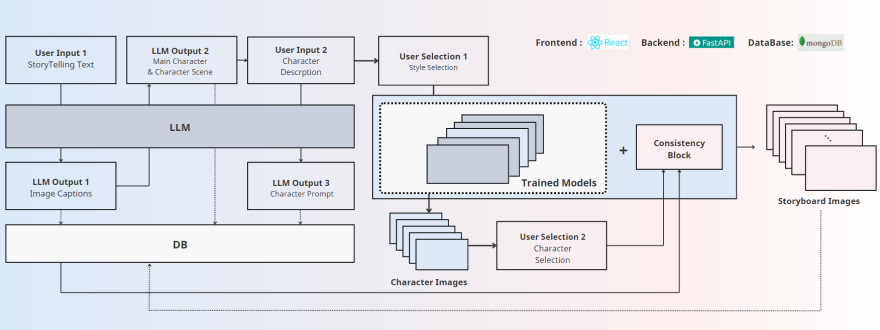

Figure 1: Pipeline of StoryWeaver

2.2 Diffusion Model

Diffusion model is a model trained by the noising and de-noising the image with multiple steps of iteration. Diffusion overcomes the limitations of GANs where (1) GANs is vulnerable to mode collapse, (2) GANs does not guarantee convergence, and (3) GANs can suffer from training when either the discriminator or the generator has too low performance. Diffusion model requires long time steps and large computation cost than GANs, however its stable and high performance drived the trend of generative models. As Diffusion model did not concentrated in the similarity between generated images, generating consistent and continuous images became one of the rising issue to solve. [5] proposed Story Diffusion model which guarantees the consistency between generated images.

3 Specific Objectives

3.1 objective

We classify our project as a text-to-image (txt2img) generation task. We generate the sequence of images which (1) guarantees consistent image generation, (2) is free from copyrights, and (3) aligns well from the given storyboard scenario text. As mentioned before, we generate image according to the storyboard scenario text. When the model generates images, the model interacts with the user for more sophisticated and better satisfaction with the usage of models. For example, the user choose the image style among the candidates, and the user inputs an image of an object which the author predefined with visual description. Our service contributes in saving time, resolving copy right issues, and maintaining the user's drawing style.

3.2 Dataset

In this project, we gathered images systematically from various online sources using web scraping techniques. We gathered the dataset by following the rule that trains the model to enforce generating various styles of images. We proprocessed the crawled data manually which does not fit to the dataset we persue, and split the images in sequence. Additionally, we divided and gathered images by the style, which are oriental, western, dessin style.

3.3 Approach

Story Weaver inputs the user prompt with storyboard scenario. LLM converts the scenario in an image capture style, and pass through the database and the LLM again. It outputs the given captioning text to describe the main character and the character scene. This text guides the user to add character description and becomes the prompt for generating sequence of images after passing the LLM. User selects the image style and a finetuned DreamBooth model with its its consistency block LoRA generates storyboard images.

Figure 2: Input image of a character

4 Final Results And Discussion

4.1 Action Recognition

Our project succeed in generating a consistent, monochrome image by using the pipeline while showing a consistent performance. 2 and 3 refers to the model input image of two characters, and 4 is the generated image maintaining the characteristics of both characters.

We have five fine-tuned DreamBooth models which are named as fantasy, sketch, casual, western, and fantastic. Each models are trained by distinguished datasets which have different image style. Western refers to the image style that we can easily view in western domain. 5 is an example image generated by western. Fantastic refers to a sketch style near to real storyboard style, where 6 is the sample image. Fantasy concentrates in more detail with the image, where 7 is the sample image. Casual persues simplicity and clean texture, which its image style is oriental. 8 refers to the sample image generated by 'casual'. Finally, sketch follows dessin or detailed sketch style following relatively realistic and western style, where 9 refers to the sample image using sketch.

5 Limitation And Future Work

5.1 Limitation

Our services includes three observed limitations. First, DreamBooth with LoRA does not guarantee the quality of the image, so some image can include inappropriate anatomy. Second, the image size can perform unstably. Third, DreamBooth is strongly dependent to the trained dataset, where the lack of data or low quality of image causes failing in the wanted image generation. Storyboard image and their similar image does not exist plentifully as people tend to not upload the incomplete or the mid-process images.

5.2 Future Work

We suggest that future work has to overcome the lack of data set, enhance the performance to reduce incomplete image, and become more robust to the generated image size.

Figure 3: Input image of a character

Figure 4: Generated with consistency between two characters

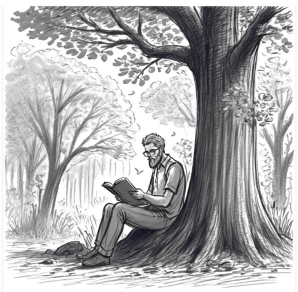

Figure 5: Sample image generated by 'western' DreamBooth model

Figure 6: Sample image generated by 'fantastic' DreamBooth model

Figure 7: Sample image generated by 'fantasy' DreamBooth model

Figure 8: Sample image generated by 'casual' DreamBooth model

Figure 9: Sample image generated by 'sketch' DreamBooth model

References

[1] 현진 박. 생성 ai 시대, "보는데 3분, 만드는데 150분 역전될까"...웹툰에 스며드는 인공지능으로 제작 환경 바꾼다. *aitimes*, 8 2023. https://www.aitimes.kr/news/articleView.html?idxno=28808.

[2] 영우 장. 웹툰작가 노동조합? "만드는 작가도 즐거운 웹툰 필요해". *aitimes*, 10 2021. https://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0002780702.

[3] 상민 성. [성상민의 문화 뒤집기] 한국 최초 '웹툰 작가 건강 조사보고서', 어떤 의미인가. *mediatoday*, 01 2023. https://www.mediatoday.co.kr/news/articleView.html?idxno=308178.

[4] 슬기 편. 네이버웹툰작가그림'ai학습'활용법사각지대놓인'저작권'. *sisaon*, 5 2023. https://www.sisaon.co.kr/news/articleView.html?idxno=150501.

[5] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024.