Just Say No to the Boring Books

Abstract

In this project, we propose a system to generate illustrations for paragraphs from the books in a specific style based on their genres to make boring novels in our literature, interesting. The pipeline of the system consists of three main parts: knowledge distillation based information retrieval model to summarize given paragraph, a clustering model to figure out the genre of the input, and fine-tuned stable diffusion model to generate image from paragraph summary given special style, which depends on genre.

Introduction

The visuals in our childhood books are some of the most memorable components, so why do we we not see them more often in adult novels as well? One may explain it as lack of style or a devotion to the force of word, or simply that advanced readers do not require illustrations. However, it is difficult to argue that an adult text with some artwork supporting the story complements a book. There are many reasons why we should use more images in the books. Images can provide additional information that supplements the text. They are helpful for communicating ideas and information, and by supporting the text's material, they can enhance understanding. They can break up large blocks of text, making the book more visually appealing. They can help set the tone and atmosphere of the book so the reader doesn't loose the attention. All these reasons compelled us to make a solution to make visualization of text paragraphs in novels to make them even more interesting. The architecture of our proposed pipeline can be seen in figure 1.

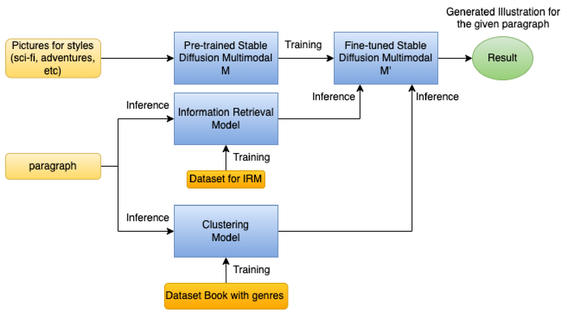


Figure 1. System Architecture

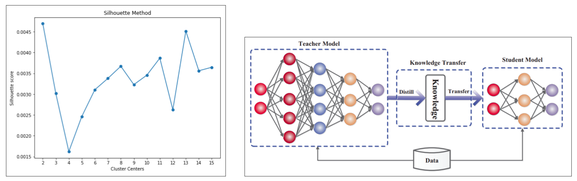
Objectives

1. Cluster all paragraphs from the books in their respective genres (adventure, detective, horror, romantic and sci-fi)
2. Train knowledge distillation based information retrieval models to extract key information from book paragraphs
3. To fine tune stable diffusion model on multiple styles from each genre to make contextual illustrations related to the provided paragraph.

Methodology

Our system consists of three main parts: Clustering, Information Retrieval and Stable Diffusion image generator.

Clustering: The clustering model is used to figure out to which genre a paragraph belongs. These genres will be used as input for the image generator model to specify which style to use. However, there is a problem with data: there are a lot of datasets with books and their genres, but there is no data with paragraphs and their genres. We can not train the model only on the whole text of books, or just assign the genre of the book to paragraph, because they can differ (for example horror books can have romantic paragraphs). We decided to create the dataset by dividing downloaded books into paragraphs and cluster all paragraphs from all books using the KMeans algorithm. We assume that paragraphs with the same genre will be in the same cluster. And then assign dominant genres of books, from where these paragraphs were obtained, to these clusters. We used 13 clusters for the KMeans algorithm because of the output of the Silhouette method as shown in figure 2(a).



(a) Silhouette Method (b) Knowledge Distillation

Figure 2. (a) Silhouette Method (b) Knowledge Distillation

Information Retrieval - Summarization: In order to generate illustrations we just need the important information from a certain paragraph. As we only had paragraphs from the books but not the labels, we used knowledge distillation approach to first make the summaries of those paragraphs and then auto-trained student summarization models with varying maximum summary lengths as shown in Figure 2(b). For the teacher models, we used **distilbart-xsum (R2: 21.37)**, **bart-large-cnn (R2: 21.06)** and **GPT3 text-davinci-002**, with **255M**, **406M** and **175B** parameters respectively to make labels (summaries) for paragraphs. Our try with GPT3 failed as it created very poor summaries, for rest of the two teacher models, we trained 3 different student models varying in maximum summary lengths **(30, 50, 100)**. Each student model is optimized one among additional 5 models, to get a total 30 combinations. The validation metrics of these top 3 (Rouge2) student model per teacher can be seen in Table 1

BART	Validation Metrics		
	Loss	Rouge1	Rouge2
Model 1	0.353	**52.79**	**46.13**
Model 2	0.286	42.38	35.02
Model 3	**0.227**	33.63	27.80

XSUM	Validation Metrics		
	Loss	Rouge1	Rouge2
Model 1	1.16	43.42	23.55
Model 2	**1.20**	**44.17**	**24.65**
Model 3	1.32	32.43	18.01

Table 1. Validation Metrics of Student Models

Stable Diffusion model fine tuning:

Data Preparation We need to collect data in order to produce a personal fine-tuned version of stable diffusion. Unlike ordinary object fine tuning, our objective is to create personalized styles that correspond to book categories. Initially, we had five book genres, which meant five unique styles. To see if the approach learns from the photos provided, we deliberately selected images that could be separated from other styles. Planets, galaxies, blue and dark blue colors, for example, can be found in the sci-fi genre. Pictures with pink backgrounds are more common in the romantic genre. etc.

Fine-tune Stable Diffusion using Textual Inversion. This approach does not change the initial stable diffusion model but creates separate embedding for the new tuned model. Which is good since it prevents us from making a mess with the model.

To begin with, we extracted all images from Google without any preprocessing, selection, or other procedures. Then, manually select images from the collected dataset because style-based fine-tuning requires a consistent color scheme and art style. Handpicked photos for style ranged from 3 to 8 pictures, as the model may not converge otherwise. Finally, the basic photos were scaled to 512x512 resolution.

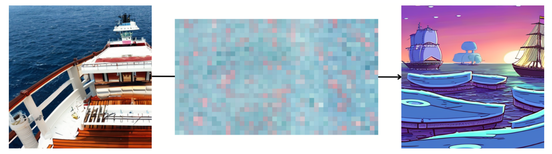
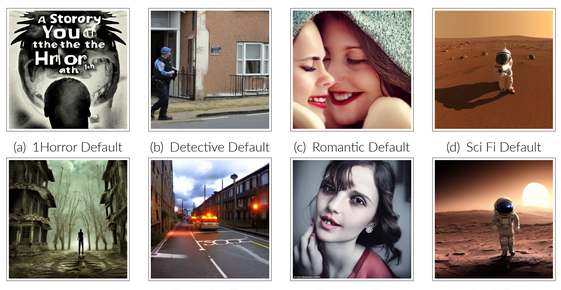


Figure 3. Caption

Results

The final illustrations created from stable diffusion for each style can be seen in figure 4. If we refer to the summaries provided in the Appendix we will know that these illustrations are quite contextual. Given a paragraph from any book our pipeline is able to classify it into certain genre and then make the summary out of it to be used by stable diffusion to make a really good illustration of the scene in a certain style.



(a) 1Horror Default (b) Detective Default (c) Romantic Default (d) Sci Fi Default

(e) Horror Tuned (f) Detective Tuned (g) Romantic Tuned (h) Sci Fi Tuned

Figure 4. Default Vs Tuned results of Stable Diffusion Model

Limitations and Future Work

- For clustering we faced some problems with certain paragraphs in adventure genre. Our model is unable to categorize paragraphs having context specifically about adventure on land. In future we plan to use contextualized embeddings instead of just vectorized ones that we used now.
- There was no proper dataset available for book paragraph summarization task. We spent most of the time in creating and cleaning our own datasets. On top of using BART and XSUM for teacher models, we plan to use T5 Model [2] to make summaries.
- Fine tuning a stable diffusion model is memory and time expensive task. We used only one sample in a batch, however with more GPU resources, we could use more samples in a batch to get better illustrations. In future we plan to use the newly release stable diffusion 2.0 for our task.

Conclusion

Even it was really difficult for the KMeans to cluster paragraphs to a specific genera we still got pretty good results. The student models trained on BART summarizer performed much better than the ones from XSUM as can be see in Table 1. Certain discoveries are worth to be mentioned. A 30 word summary delivers more accurate visuals, but the text material loses quality. Although the 100-word description contains the greatest information, the stable diffusion model produces more abstract and non-obvious pictures. So 50 words is the appropriate number of words.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[2] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Appendix

Text that was used to produce images.

- She knew that feeling and knew its signs, and saw them in Anna; saw the quivering, flashing light in her eyes, and the smile of happiness and excitement unconsciously playing on her lips
- The story of a man who has been working with a young man in his head.
- I wasn't expecting to be first at anything. I was the 5th crewman out of the MDV when we landed, making me the 17th person to set foot on Mars
- Police considered that it might be caused by Arthur's closing his bedroom door.
- The story of a ship that was on the deck at the end of the year.