Story to Illustration: Automated Literature Illustration Leveraging Pre-trained Diffusion Model

Torstein Ørbeck Eliassen, Yuntao Ma and Asta Wu {torsteoe, yma42, astawu}@stanford.edu

Stanford University

Problem

Illustrations of stories can increase accessibility and enrich a reader's understanding. We propose a method for automating the illustration process by leveraging the open-source Stable Diffusion model [1]. We address two key challenges: 1) **summarizing passages** from well-known stories for prompt generation via sequence to sequence (seq2seq) models and 2) **maintaining stylistic consistency** between generated images via neural style transfer (NST).

Dataset & Features

Original Stories: Daedalus & Icarus, Echo & Narcissus, and H.C. Andersen: The Little Match Girl

Text summarization Task: CNN/Daily Mail with 311,971 news articles and highlights

Models

Text Summarization: Seq2seq transformer model pretrained on BART[2] and finetuned on CNN/Daily Mail

Image Generation: Pretrained Stable Diffusion

Neural Style Transfer[3]: Using a pretrained VGG and backpropagating on inputs to push Gram matrices of activations together. Pretrained ResNet50 for testing.

Results & Discussion

Text Summarization:

ROUGE-N scores are based on the no. of matching n-gram between the generated text and the reference text, with higher scores indicating more similarity.

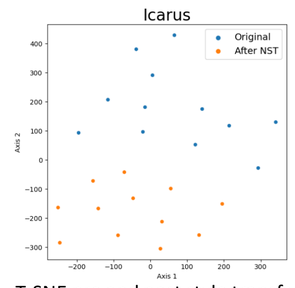
Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
LSA	unsupervised	20.48	5.03	17.85
B-LSTM	CNN/Daily Mail	27.93	13.32	18.34
BART	Xsum	24.27	7.03	16.99
BART	CNN/Daily Mail	41.957	20.76	30.597

Style consistency:

- Considerable decrease in standard deviation of Gram matrices on both VGG19 and ResNet50
- T-SNE on Gram matrices visualizes distinct clusters

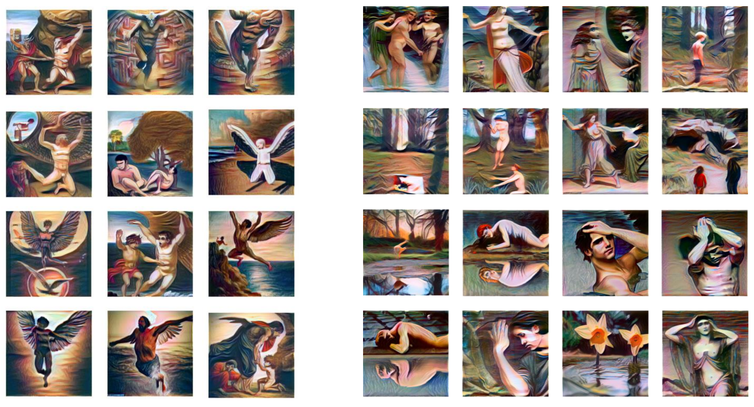
Story	VGG ("train")	ResNet ("test")
Icarus (w. ref)	80.3%	19.9%
Narcissus (w. ref)	83.0%	24.8%
LMG (w.ref)	90.03%	28.8%
Icarus (0 ref)	93.97%	19.11%
Narcissus (0 ref)	94.5%	22.6 %
LMG (0 ref)	94.5%	19.7%

Style disparity decrease



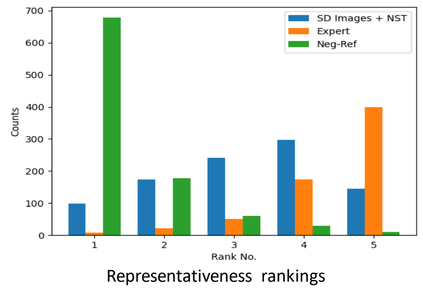
T-SNE pre and post style transfer

Final images: Icarus & Narcissus



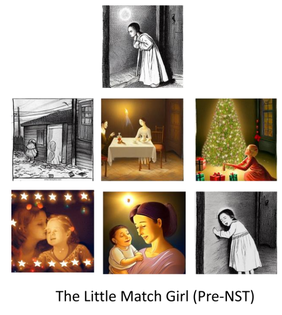
Human Evaluation:

- Our illustrations were average in representativeness relative to expert and negative reference photos.



Representativeness rankings
(1 - not representative 5 - represents passage well)

- 84% (208 / 246) of evaluators voted the images more stylistically consistent after NST
- Average content preservation rank is 4.35 / 5 (5 is well-preserved) for pre- and post- NST images.



The Little Match Girl (Pre-NST)

The Little Match Girl (Post-NST)

Future Work

- Additional work on prompt to capture the larger story context in each prompt.
- Maintaining other forms of consistency such as character consistency across illustrations.

References

[1] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," 2022.
[2] M. Lewis et al., "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019
[3] L. A. Gatys et al., "A neural algorithm of artistic style," 2015