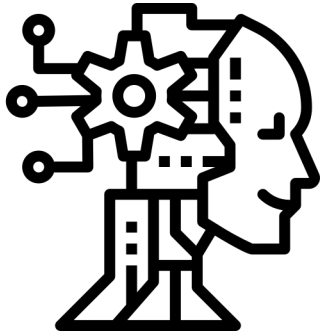


FairyTailor

Andoni Garrido, Giles Desmidt and Iñigo Auzmendi

INDEX

WHAT IS FAIRYTAILOR?

- A model for generating storytellings with text and images.
- Made by IBM researchers.
- Two different uses:
  - Generate all the story using giving a title.
  - Cooperate with writers.
- One prototype have been released.

DATASET USED

- Text dataset:
  - Reddit's WritersPrompt
  - Manually made dataset of children's books.

  300.000 stories.

- Images:
  - Flicker30

  31.000 images.

PROMPT:
Every year, the richest person in America is declared the "Winner of Capitalism". They get a badge, and all of their wealth is donated to charity, so they have to start back up at $0

STORY CREATED BY 'Damptruff1':

The CEO sat in his office. It had a deep red for a carpet, and quite a few coffee stains. The walls were painted a beautiful white, with his desk and the cabinets made out of a wood with a rich brown. He himself wore a gray suit, with a red tie and a white undershirt. He preferred a sweater and sweatpants, but today was an important meeting...

SYSTEM ARCHITECTURE

Two architectures attempts:

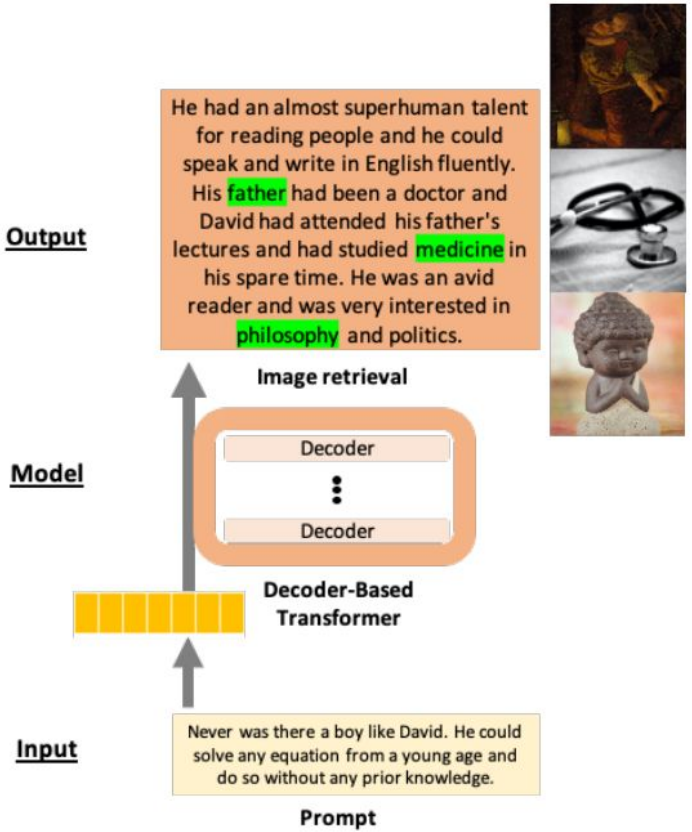
1. Benchmark model
2. Final Design

Benchmark Design

- **Generates text** and accordingly **retrieves images.**
- **Tests** readability, diversity and sentiment of the generated text.

TEXT	IMAGES

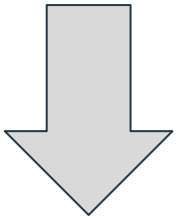
- Taken **GPT-2 model two fine-tuning** were made:
  - **Reddit WritingPrompt** [Fan et al., 2018] to fine-tune the model **to a prompt-story template.**
  - **Adapt** the model **based on** individually collected **children's books dataset.**
- Tested model ***top-k random sampling*** **method (with k=50)** used in the *Hierarchical Neural Story Geneation* model (with k=10) [Fen et al., 2018]
	- This **architecture extracts most common nouns from the generated text**
	- Then, **retrieve** the corresponding **images from Flickr30K** [Plummer et al., 2017]



Benchmark architecture scheme

Limitations

- Text completions are often repetitive, incoherent, inappropriate and dark.
- The independently retrieved images are inconsistent



IMPROVEMENT

Final Design - Text modality

- Several re-ranker metrics added to increase the readability, positiviness, coherency and tale-like manner.
- The re-rankers computes the minmax (1) normalization to rescale each feature across all generated texts so that all features contribute equally.

$$scaled\_scores = \frac{scores - \min(scores)}{\max(scores) - scores} \quad (1)$$

Final design - Text modality

- The r**e-ranker frequency has been increased** to obtain a coherent text generation, **Re-rank after each end-of-sentence token**
- Features taken into account in the re-ranks:
  - Readability
  - Positive Sentiment
  - Diversity
  - Simplicity
  - Coherency
  - Tale-like

Re-rank features

Readability

Calculates the length of sentences and length of words to estimate how complex the text is.

*readability = 0.5 * word_chars + sent_words*

Positive Sentiment

- **SentiWordnet** [Baccianella et al., 2010] to compute positivity polarity, assign sentiment scores to each **WordNet** synonym group.
- **WordNet** is popular for in**formation retrieval tasks** and **does not require pre-training.**

Re-rank features

Simplicity

Calculates the fraction of tale-like characteristic words in the given text.

$$simplicity = \text{len}(\text{set}(filtered\_words) \cap freq\_words)$$

Diversity

Calculates the fraction of unique words from the total number of words.

$$diversity = \frac{\text{len}(\text{set}(filtered\_words))}{\text{len}(filtered\_words)}$$

Re-rank features

Coherency

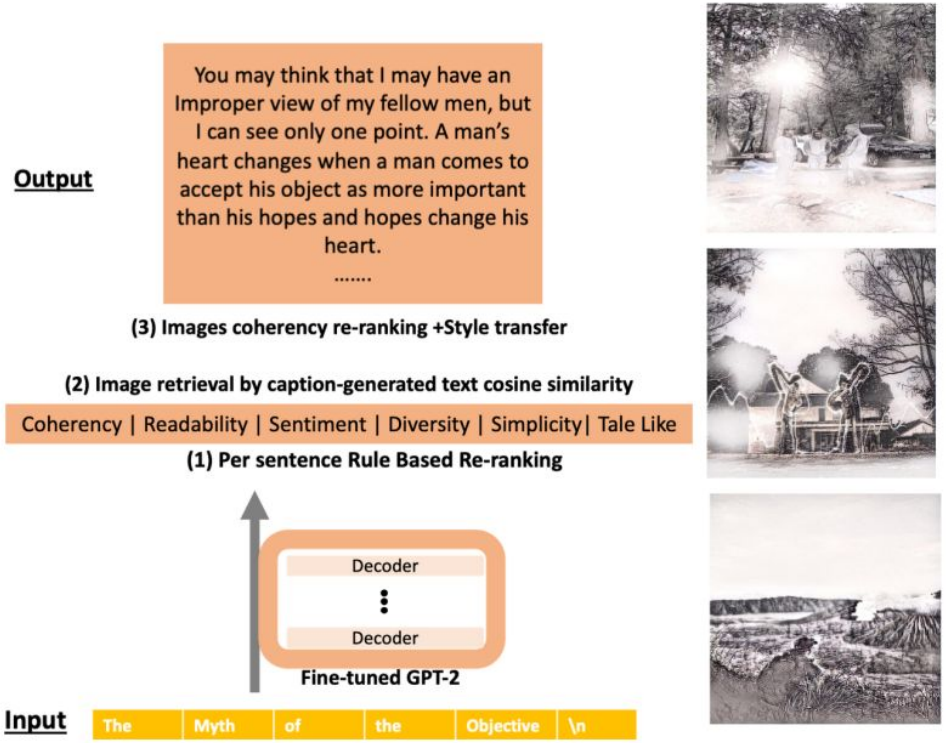
Calculates the Latent Semantic Analysis (LSA) similarity within the story sentences compared to the first sentence.

Tale-like

Tale like computes the KL divergence loss between a preset GPT-2 and a fine-tuned GPT-2 generated texts' prediction scores.

Image Modality

- **Three open-source implementations** for text to **image synthesis** are evaluated:
  - BigGAN [Brock et al., 2018],
  - stackGAN [Zhang et al., 2017]
  - Dall-E [Ramesh et al.,2021]
- Image generation 4-30 s **VS** Image retrieval 0.5-2 s
- To **compute the similarity between text and images**, the **cosine similarity** of the text embeddings and the images embeddings are computed.
  - RETURN: the images' ids of the highest-scoring images.
- The images' consistency metric is calculated.



Final design scheme

Prototype “FairyTailor”

- Multimodal framework allowing story co-creation
- Human writer starts in multiple ways
- FairyTailor autocompletes the rest
- Storytelling on a new level
- Perfect mix between human and machine
- **DEMO**

Conclusion

- Well advanced technique
- Constantly developing
- Used more and more
- Storytelling on a new level
- Perfect mix between human and machine

REFERENCES

- **Eden Bensaid, Mauro Martino, Benjamin Hoover, Hendrik Strobelt [FairyTailor: A Multimodal Generative Framework for Storytelling] 2021**
  https://arxiv.org/abs/2108.04324

- **Eden Bensaid's github**. https://github.com/EdenBD/MultiModalStory-demo

- **FairyTailor prototype.** https://fairytailor.org/

- **Reddit's WrittingPrompts thread.** https://www.reddit.com/r/WritingPrompts/

- **Sonali Fotedar , Koen Vannisselroij , Shama Khalil , Bas Ploeg [Storytelling AI: A Generative Approach to Story Narration] 2020**
  http://ceur-ws.org/Vol-2794/paper4.pdf

- **Radford, J.Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2018. URL** https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
- **T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers:** State-of-the-art natural language processing. ArXiv, abs/1910.03771, 2019.
- **A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation, 2018.**
- **A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration, 2019.**
- **A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities:** Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV, 123(1):74–93, 2017.
- **Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, LREC. European Language Resources Association, 2010. ISBN 2-9517408-6-7.** URL http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf

- Fellbaum. WordNet: An Electronic Lexical Database. Bradford Books, 1998.
- Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis, 2018.
- Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017.
- Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation, 2021.
- Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, 2016.
- He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

THANK YOU

Iñigo Auzmendi, Guiles Desmidt and Andoni Garrido