STORY GENERATION WITH UNSUPERVISED PRE-TRAINING

A comparative analysis of Supervised & Unsupervised Pre-Training...

Final Project Presentation By:
Lalit Gupta[2021201018], Sourabh Patidar[2021201089], Shankar Jetty Nag[2021900019], Aaryan Singh[2019114017]

Github Repository

INTRODUCTION

Story generation has become an increasingly popular research AREA in natural language processing (NLP), with many applications in areas such as creative writing, automated dialogue systems, and video game design.

PRE-TRAINED models can help improve the performance of story generation TASK BY providing them with a better understanding of natural language and the ability to generate more coherent and engaging narratives.

WHAT IS TO EXPLORE

The aim of the Project is to compare the effectiveness of supervised and unsupervised pre-training approaches for story generation.

Specifically, we aim to investigate how these approaches differ in terms of their ability to generate engaging and coherent stories and whether one approach is more effective than the other.

RESEARCH PROBLEMS

- What are the differences in performance between supervised and unsupervised pre-training approaches for story generation?
- How do these approaches compare in terms of their ability to generate stories with different levels of complexity and coherence?

METHODOLOGY

DATASETS

PRE-TRAINED MODELS

FINE-TUNING FOR STORY GENERATION

EVALUATION

ANALYSIS

DATASET-

ROC-Stories:

- We have used the ROC Stories dataset for our experiments, which consists of a collection of 5 - sentence stories. The dataset contains a total of 1,91,413 stories.
- We used the standard train and validation splits of the dataset, with 1,76,688 (92.5%) stories for training and 9,816(5%) stories for validation. We held out the remaining 4,909(2.5%) stories for testing the performance of our models.

PRE-TRAINED MODELS

- We used GPT-2 as our unsupervised pre-training approach, using the Hugging Face Transformers library. We fine-tuned the model on the ROC Stories dataset for 44172 steps using the default hyperparameters.
- While for the supervised approach, we have used MVP-Story from MVP (Multi-task Supervised Pre-training for Natural Language Generation) . MVP-story is a prompt-based model that MVP is further equipped with prompts pre-trained using labeled story generation datasets.

FINE-TUNING FOR STORY GENERATION

- We have fine-tuned the gpt-2 model on ROC-Stories dataset.
- We have fine-tuned the model with 44172 steps and 1 epoch and kept other hyperparameters as default.

```
***** Running training *****
  Total_num_training_step = 44172
  Num Epochs = 1
  Train_batch_size per device = 4
  Valid_batch_size per device = 4
Start epoch1 of 1
(batch loss=2.60208): 100%|██████████| 44172/44172 [1:48:27<00:00,  6.79it/s]
Average train loss per example=2.6197360202927826 in epoch1
Starting evaluate after epoch 1
eval: 100%|██████████| 2454/2454 [01:27<00:00, 28.13it/s]
Average valid loss per example=2.4752597598011357 in epoch1
Perplextiy for valid dataset in epoch1 is 11.884793903870971
```

MERICS USED

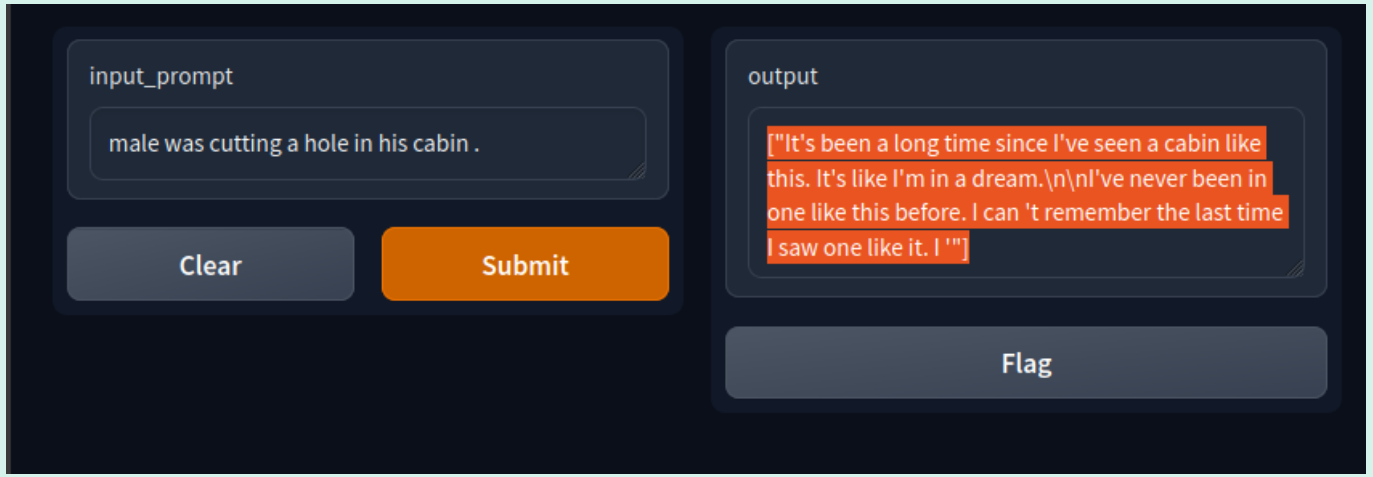
- We used perplexity and BLEU score as quantitative metrics to measure the performance of our models.
- Perplexity measures how well the model has adapted to the natural language, while BLEU score measures how well the generated story matches the reference story.
- We also have used Distinct Metric which gives us the diversity score of generated text.

RESULTS

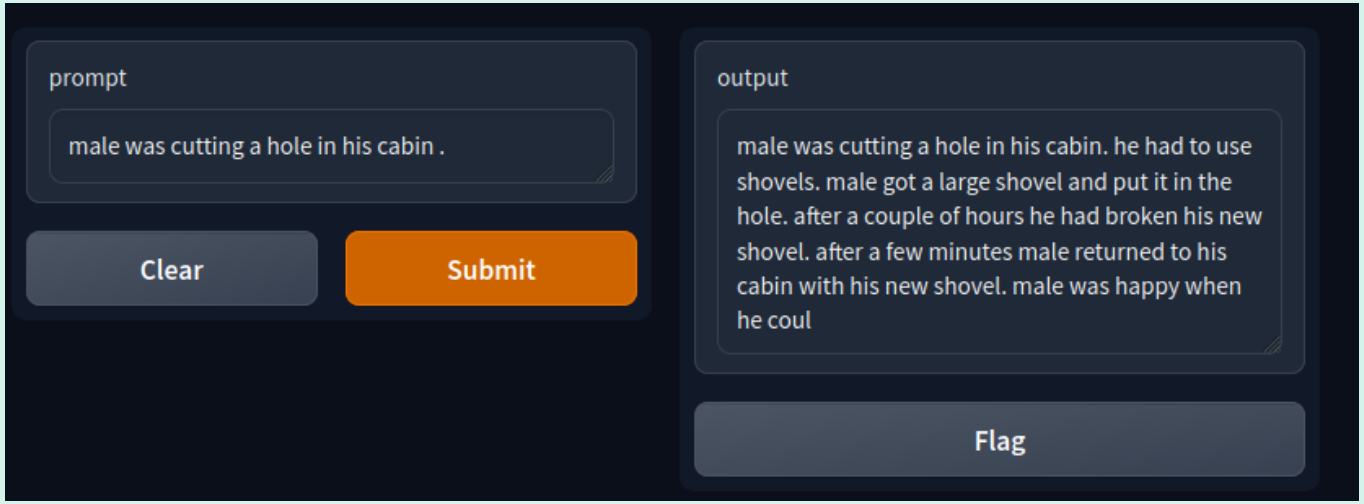
	MVP-STORY*	GPT-2
BLEU-1	33.79	6.2
BLEU-2	15.76	0.0005
DISTINCT-1	3.02	5.53
DISTINCT-2	75.65	34.16

*- These results are taken from the MVP (Multi-task Supervised Pre-training for Natural Language Generation) paper.

SAMPLE RESULTS

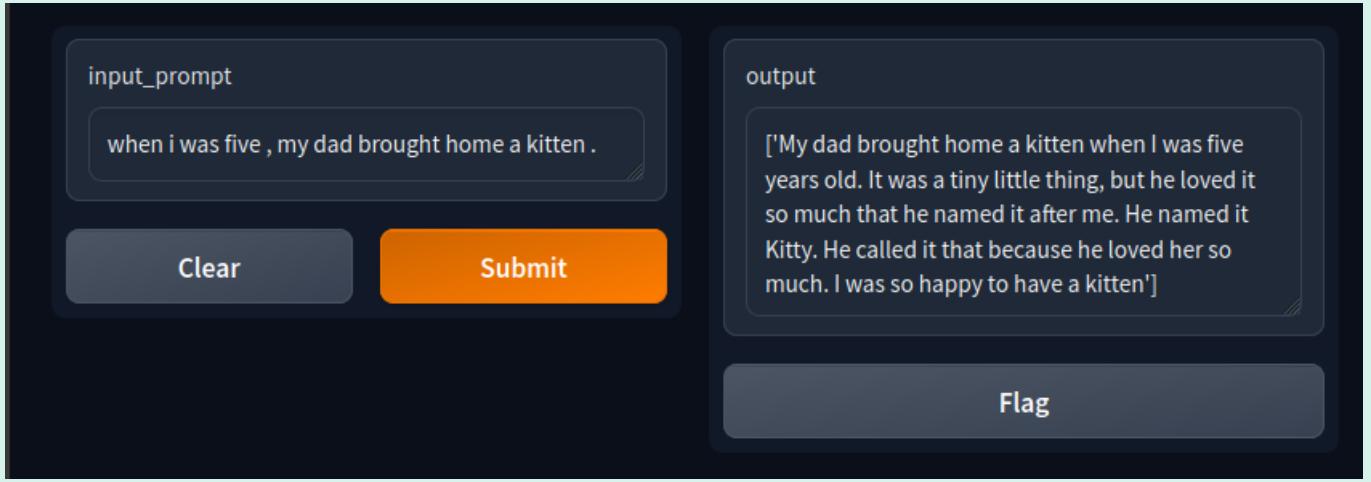


MVP-STORY MODEL(SUPERVISED)

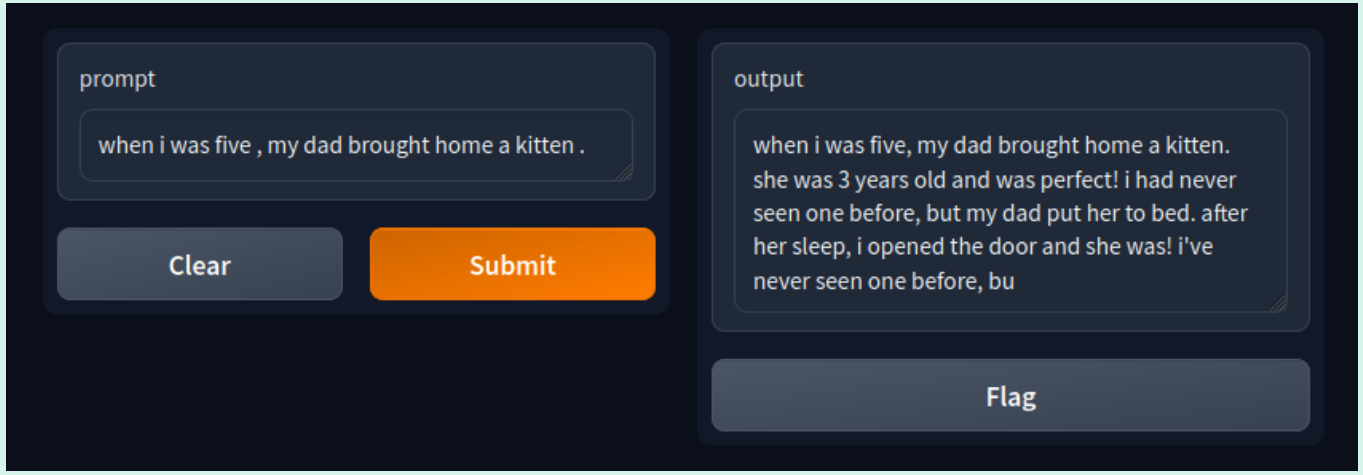


GPT-2 MODEL(UNSUPERVISED)

SAMPLE RESULTS



MVP-STORY MODEL(SUPERVISED)



GPT-2 MODEL(UNSUPERVISED)

LIMITATIONS

- Supervised pre-training, which may not be representative of all possible story types limiting the genres in which stories are being generated in good quality.
- Additionally, our evaluation metrics may not fully capture all aspects of story quality, we only have one reference story to calculate the BLEU score. Hence, we cannot conclude the story generated by the model is very bad.

KEY TAKEAWAYS

- Supervised Pre-training is giving better results in terms of story generation than unsupervised pre-training, but this also creates a problem when it comes to build huge structured datasets containing diversity of prompt to story.
- As the MVP paper argues that multi-task pre-training can lead to better generalization as the pre-trained model learns more diverse and robust representations of language.

THANK YOU