simple-prompt-analysis

November 5, 2022

1 ML PROJECT - RESEARCH PAPER IMPLEMENTATION

2 TOPIC - DIFFUSIONDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models

3 PROMPT ANALYSIS

3.0.1 GROUP MEMBERS:

3.0.2 RITU MEWADA(191310132003)

3.0.3 DRASHTI VAGHELA (191310132014)

3.0.4 DHRUVA VAIDYA (191310132015)

3.0.5 JINAL VYAS(191310132019)

Setup the environment, if needed

[2]:
```
## Update the following with your specific version of CUDA, if any.
!pip install torch --extra-index-url https://download.pytorch.org/whl/cu113
!pip install h5py pandas numpy matplotlib diffusers transformers scipy ftfy␣
↪pyarrow regex wordcloud
```

```
Collecting h5py
  Using cached
h5py-3.7.0-cp39-cp39-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (4.5 MB)
Collecting pandas
  Using cached
pandas-1.5.1-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.2 MB)
Collecting numpy
  Using cached
numpy-1.23.4-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (17.1 MB)
Collecting matplotlib
  Using cached
matplotlib-3.6.1-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (11.8
```

```
MB)
Collecting diffusers
  Using cached diffusers-0.6.0-py3-none-any.whl (255 kB)
Collecting transformers
  Using cached transformers-4.23.1-py3-none-any.whl (5.3 MB)
Collecting scipy
  Using cached
scipy-1.9.3-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (33.8 MB)
Collecting ftfy
  Using cached ftfy-6.1.1-py3-none-any.whl (53 kB)
Collecting pyarrow
  Downloading
pyarrow-10.0.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (35.2
MB)
                            35.2/35.2 MB
30.3 MB/s eta 0:00:0000:0100:01
Collecting regex
  Downloading
regex-2022.10.31-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (769
kB)
                           770.0/770.0 kB
72.1 MB/s eta 0:00:00
Collecting wordcloud
  Using cached
wordcloud-1.8.2.2-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (458
kB)
Requirement already satisfied: python-dateutil>=2.8.1 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from pandas)
(2.8.2)
Collecting pytz>=2020.1
  Using cached pytz-2022.5-py2.py3-none-any.whl (500 kB)
Collecting kiwisolver>=1.0.1
  Using cached
kiwisolver-1.4.4-cp39-cp39-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (1.6
MB)
Requirement already satisfied: pyparsing>=2.2.1 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from
matplotlib) (3.0.9)
Requirement already satisfied: packaging>=20.0 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from
matplotlib) (21.3)
Collecting pillow>=6.2.0
  Downloading
Pillow-9.3.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.2 MB)
                            3.2/3.2 MB
54.3 MB/s eta 0:00:00a 0:00:01
Collecting cycler>=0.10
  Using cached cycler-0.11.0-py3-none-any.whl (6.4 kB)
```

```
Collecting contourpy>=1.0.1
  Downloading
contourpy-1.0.6-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (296
kB)
                           296.3/296.3 kB
55.3 MB/s eta 0:00:00
Collecting fonttools>=4.22.0
  Using cached fonttools-4.38.0-py3-none-any.whl (965 kB)
Collecting filelock
  Using cached filelock-3.8.0-py3-none-any.whl (10 kB)
Collecting requests
  Using cached requests-2.28.1-py3-none-any.whl (62 kB)
Requirement already satisfied: importlib-metadata in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from
diffusers) (5.0.0)
Collecting huggingface-hub>=0.10.0
  Using cached huggingface_hub-0.10.1-py3-none-any.whl (163 kB)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1
  Using cached
tokenizers-0.13.1-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (7.6
MB)
Collecting pyyaml>=5.1
  Using cached PyYAML-6.0-cp39-cp39-manylinux_2_5_x86_64.manylinux1_x86_64.manyl
inux_2_12_x86_64.manylinux2010_x86_64.whl (661 kB)
Collecting tqdm>=4.27
  Using cached tqdm-4.64.1-py2.py3-none-any.whl (78 kB)
Requirement already satisfied: wcwidth>=0.2.5 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from ftfy)
(0.2.5)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from
huggingface-hub>=0.10.0->diffusers) (4.4.0)
Requirement already satisfied: six>=1.5 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from python-
dateutil>=2.8.1->pandas) (1.16.0)
Requirement already satisfied: zipp>=0.5 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from
importlib-metadata->diffusers) (3.10.0)
Collecting urllib3<1.27,>=1.21.1
  Using cached urllib3-1.26.12-py2.py3-none-any.whl (140 kB)
Collecting charset-normalizer<3,>=2
  Using cached charset_normalizer-2.1.1-py3-none-any.whl (39 kB)
Requirement already satisfied: idna<4,>=2.5 in
/nvmescratch/hoo/miniconda3/envs/dummy/lib/python3.9/site-packages (from
requests->diffusers) (3.4)
Collecting certifi>=2017.4.17
  Using cached certifi-2022.9.24-py3-none-any.whl (161 kB)
Installing collected packages: tokenizers, pytz, urllib3, tqdm, regex, pyyaml,
```

```
pillow, numpy, kiwisolver, ftfy, fonttools, filelock, cycler, charset-
normalizer, certifi, scipy, requests, pyarrow, pandas, h5py, contourpy,
matplotlib, huggingface-hub, wordcloud, transformers, diffusers
Successfully installed certifi-2022.9.24 charset-normalizer-2.1.1
contourpy-1.0.6 cycler-0.11.0 diffusers-0.6.0 filelock-3.8.0 fonttools-4.38.0
ftfy-6.1.1 h5py-3.7.0 huggingface-hub-0.10.1 kiwisolver-1.4.4 matplotlib-3.6.1
numpy-1.23.4 pandas-1.5.1 pillow-9.3.0 pyarrow-10.0.0 pytz-2022.5 pyyaml-6.0
regex-2022.10.31 requests-2.28.1 scipy-1.9.3 tokenizers-0.13.1 tqdm-4.64.1
transformers-4.23.1 urllib3-1.26.12 wordcloud-1.8.2.2
```

4 Analyzing prompts provided by DiffusionDB

[3]:
```
from PIL import Image
from pathlib import Path
import os
import json
from diffusers import StableDiffusionPipeline
import regex as re
import pandas as pd
from tqdm.auto import tqdm
import numpy as np
import matplotlib.pyplot as plt
import wordcloud as wc
import requests
```

```
/nethome/bhoover30/miniconda3/envs/dummy/lib/python3.9/site-
packages/tqdm/auto.py:22: TqdmWarning: IProgress not found. Please update
jupyter and ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
```

Load the pipeline to get the same tokenizer used as Stable Diffusion

[5]:
```
path_to_prompt_parquet = "https://huggingface.co/datasets/poloclub/diffusiondb/
 ↪resolve/main/metadata.parquet"
prompts = pd.read_parquet(
    path_to_prompt_parquet,
    columns=['prompt']
)
print("Length of prompts: ", len(prompts))
```

```
Length of prompts:  2000000
```

4.1 Prompt uniqueness?

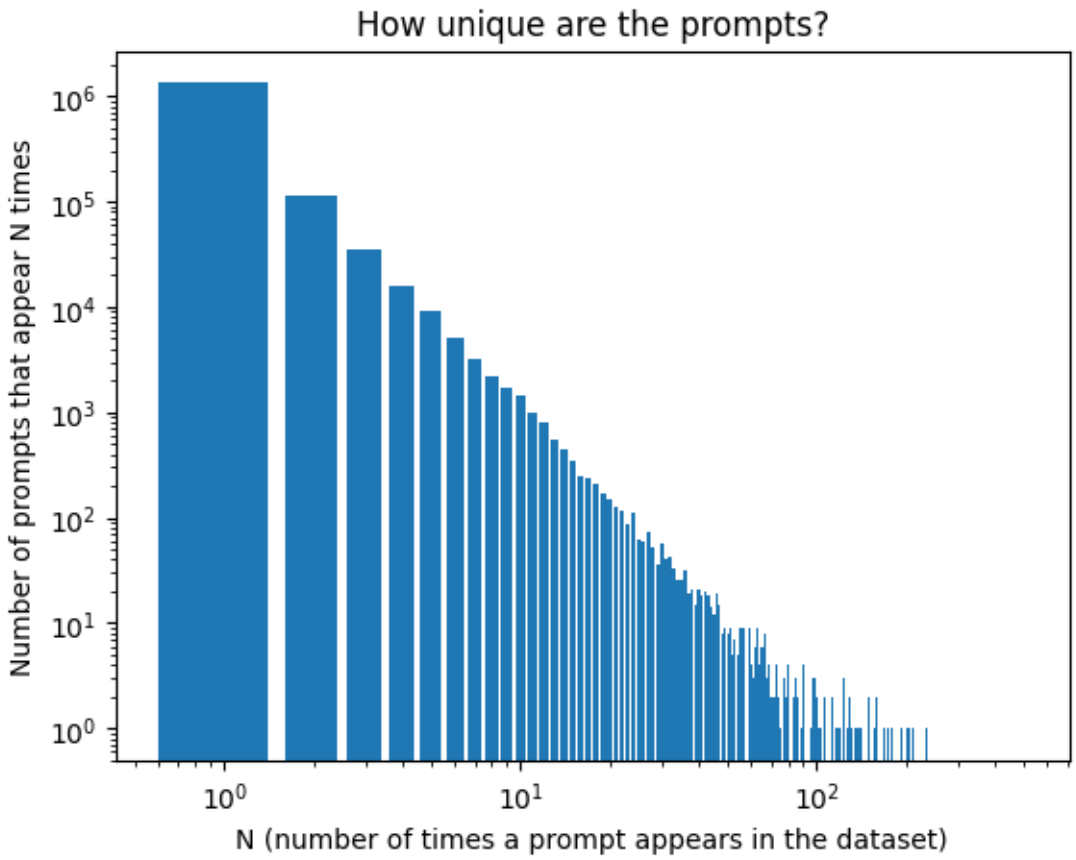
[6]:
```
sprompts = set(list(prompts.prompt))

# Get count of each prompt
ct_dict = {k:0 for k in sprompts}
for k in prompts.prompt:
    ct_dict[k] += 1
```

[7]:
```
x = np.array([v for v in ct_dict.values()])
cts, bins = np.histogram(x, bins=np.unique(x))

plt.bar(bins[:-1], cts)
plt.yscale("log")
plt.xscale("log")
plt.xlabel("N (number of times a prompt appears in the dataset)")
plt.ylabel("Number of prompts that appear N times")
plt.title("How unique are the prompts?")
```

[7]:
```
Text(0.5, 1.0, 'How unique are the prompts?')
```



4.2 Concept Frequency

A qualitative analysis of the concepts present in DiffusionDB. We manually filter the top tokens for stop words, combining subtoken representations into meaningful concepts, before displaying in a WordCloud.

[4]:
```
auth_token = os.environ["HFTOKEN"]
pipe = StableDiffusionPipeline.from_pretrained("CompVis/stable-diffusion-v1-4",␣
↪use_auth_token=auth_token)
```

```
Fetching 16 files: 100%|          | 16/16 [00:00<00:00, 14211.96it/s]
```

[22]:
```
# Show top K tokens in the corpus, visually filter as needed
cts, idxs = tokfreqs.topk(k=100)
print("\n".join([" :: ".join((str(pipe.tokenizer._convert_id_to_token(idx.
↪item())), str(cts[i].item()))) for i, idx in enumerate(idxs)]))
```

```
<|endoftext|> :: 83099494
,</w> :: 10847176
<|startoftext|> :: 2000000
a</w> :: 1523457
of</w> :: 1296773
and</w> :: 1157827
by</w> :: 1110022
art</w> :: 784355
in</w> :: 756606
detailed</w> :: 714959
the</w> :: 709072
art :: 631981
.</w> :: 613519
on</w> :: 495852
station</w> :: 476150
painting</w> :: 438349
k</w> :: 429432
with</w> :: 425057
portrait</w> :: 399555
realistic</w> :: 365993
-</w> :: 342329
8</w> :: 323039
highly</w> :: 319087
lighting</w> :: 310602
digital</w> :: 295669
intricate</w> :: 276934
beautiful</w> :: 276268
concept</w> :: 256254
trending</w> :: 245511
```

greg</w> :: 245421
0</w> :: 239802
style</w> :: 235599
4</w> :: 235164
cinematic</w> :: 229357
sharp</w> :: 228603
rut :: 225678
kowski</w> :: 222008
render</w> :: 221661
illustration</w> :: 221422
an</w> :: 216964
focus</w> :: 210662
high</w> :: 188288
fantasy</w> :: 177511
octane</w> :: 176801
m</w> :: 172288
1</w> :: 170219
d</w> :: 167528
ger :: 166860
face</w> :: 162641
photo</w> :: 161967
light</w> :: 155787
3</w> :: 155322
5</w> :: 146905
from</w> :: 146107
as</w> :: 144378
2</w> :: 132065
black</w> :: 131100
wearing</w> :: 130106
dark</w> :: 124368
smooth</w> :: 120759
white</w> :: 119682
hyper :: 117479
very</w> :: 116089
engine</w> :: 115067
unreal</w> :: 114896
background</w> :: 114650
elegant</w> :: 111326
9</w> :: 110904
hair</w> :: 110355
full</w> :: 109023
hyper</w> :: 107780
mucha</w> :: 105940
photo :: 105439
at</w> :: 102008
se</w> :: 99516
shot</w> :: 98968
woman</w> :: 96757

```
body</w> :: 96685
ultra</w> :: 96497
oil</w> :: 95192
red</w> :: 95030
alphon :: 95022
detail</w> :: 94338
is</w> :: 93762
colors</w> :: 93602
hd</w> :: 92955
eyes</w> :: 92851
(</w> :: 91822
girl</w> :: 88013
masterpiece</w> :: 87233
anime</w> :: 86647
dramatic</w> :: 86235
man</w> :: 85231
character</w> :: 84678
studio</w> :: 84033
epic</w> :: 82639
to</w> :: 81676
metric</w> :: 80573
w :: 80495
photography</w> :: 77450
```

[24]:
```
# Filtered and combined tokens
words = {
  "art": 784355,
  "detailed": 714959,
  "artstation": 476150,
  "painting": 438349,
  "portrait": 399555,
  "realistic": 365993,
  "8k": 323039,
  "highly": 319087,
  "lighting": 310602,
  "digital": 295669,
  "intricate": 276934,
  "beautiful": 276268,
  "concept": 256254,
  "trending": 245511,
  "style": 235599,
  "4k": 235164,
  "cinematic": 229357,
  "sharp": 228603,
  "greg rutkowski": 222008,
  "render": 221661,
  "illustration": 221422,
```

```
  "focus": 210662,
  "high": 188288,
  "fantasy": 177511,
  "octane": 176801,
  "face":162641,
  "photo":161967,
  "light": 155787,
  "black": 131100,
  "wearing": 130106,
  "dark": 124368,
  "smooth": 120759,
  "white": 119682,
  "hyper": 117479,
  "unreal engine": 114896,
  "background": 114650,
  "elegant": 111326,
  "hair": 110355,
  "full": 109023,
  "mucha": 105940,
  "hyper": 107780,
}

print(len(words))
```

```
40
```

[25]:
```
cloud = wc.WordCloud(width=450, height=300, background_color="white",␣
↪min_font_size=10, relative_scaling=0.0001, colormap="Dark2").fit_words(words)
im = cloud.to_image()
im
```

[25]:


painting
cinematic
intricate
8k
wearing
greg rutkowski
full
portrait
art
trending
highly
unreal engine
background
elegant
lighting
photo
4k
hyper
focus
render
smooth
artstation
fantasy
black
face
style
sharp
beautiful
dark
hair
detailed
high
light
white
mucha
octane
concept
realistic
digital
illustration