Motivation

Deep Embeddings

Representing an image by continuous embeddings from deep networks.
Motivation

Deep Embeddings vs. Quantized Tokens

Representing an image by quantized tokens from deep networks.

VQ-VAE, 2017
DALL-E, 2021
Parti, 2022

......
Deep Embeddings vs. Quantized Tokens vs. Text

Can we represent an image as text?

encoder

decoder

visual representation

a photograph closeup dog corgi wearing white hat corgi yellow glasses atop on red handle bicycle street nyc corgi smiling smiling wearing an orange wearing fedora with smile yellow sunglasses people arm among people car multiple screen buildings and street buildings cute teeth ……
De-Diffusion Text

an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street
an attributing provided closeup dog corgi wearing white yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing handle fedora with smile yellow sunglasses yellow sunglasses and people and street with foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street
Example

De-Diffusion Text

an attributing provided closeup dog corgi wearing white yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street with foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street
Example

De-Diffusion Text

an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street with foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street
an attributing provided closeup dog corgi wearing white yellow hat
corgi and hat on glasses aboard an on red handle cart street nyc
oscanadian corgi corgi presented description called an dog dog
shown standing smiling incorporating wearing an wearing
harnehandle fedora with smile yellow sunglasses yellow sunglasses
and people and street with foreground with people people among
people cars multiple billboards cityscape and street billboards
asphalt street cute screenshot street
Example

De-Diffusion Text

an attributing provided closeup dog corgi wearing white yellow hat
corgi and hat on glasses aboard an on red handle cart street nyc
oscanadian corgi corgi presented description called an dog dog
shown standing smiling incorporating wearing an wearing
handle fedora with smile yellow sunglasses yellow sunglasses
and people and street with foreground with people people among
people cars multiple billboards cityscape and street billboards
asphalt street cute screenshot street
Example

De-Diffusion Text

an attributing provided closeup dog corgi wearing white yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street with foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street
Example

De-Diffusion Text

an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing handled fedora with smile yellow sunglasses and people and street with foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street
Humans are used to encode many different modalities into text.
Humans are used to encode many different modalities into text.
“In daily life, language often acts as an interface to higher-level cognition.”
Why not?

Text as Cross-Modal Representation

Why not captions?
- Captions are usually not as comprehensive.

Why not deep embeddings?
- Interpretability
- Text can be directly used in LLMs & text-to-img models, no need of model tuning
Method

An Autoencoder

- Training only required images.
**Method**

**Diffusion as Decoder**

- Training only required images.
- Unpacking the knowledge encapsulated within the text-to-image generative models.
**Method**

**Diffusion as Decoder**

- Training only required images.
- Unpacking the knowledge encapsulated within the text-to-image generative models.
Method

**Diffusion as Decoder**

- Training only required images.
- Unpacking the knowledge encapsulated within the text-to-image generative models.
  - To minimize the reconstruction error, the text has to be comprehensive.
Method

Full Picture

- Training only required images.
- Unpacking the knowledge encapsulated within the text-to-image generative models.
  - To minimize the reconstruction error, the text has to be accurate and comprehensive.
- Discrete text tokens, trained with gumbel softmax.
Application # 1

Transferable Text -to- Image Prompt

original

Imagen

Midjourney

Stable Diffusion XL

an d avilishlis blog closeup berries jar through large refrerefre jar glass jar each other glass on an on peach hardwood closeup glass homemade mixed glass jar called relating called an oranges fruit shown slices each other containing relating an orange orange slices slices between black grapes open chunks orange oranges and berry black blueberry consist though towards pink closeup facing that background pink wall background pink pink wall wall closeup chia grapes recipe
Applicatıon #1

Compare with Human Annotation

original

Imagen

Midjourney

Stable Diffusion XL

[Human-annotated caption] A jar filled with different types of fruit on a table.

Why is this different from captions?
- Captions are usually not as comprehensive.
Benchmark on a Third-Party Diffusion Model

FID (↓): Similarity measure
Diffusion Model: Stable Diffusion v2
Dataset: MS-COCO Val 2014 30K
Application #1

Prompt for Synthetic Images

original

Imagen

Midjourney

Stable Diffusion XL

an art rha digitally s art illustration woman face wearing colorful colorful paints face painted head pink lipstick though an among colourful confetti confetti realism pin up os janu monroe monroe resembrelating called an face woman shown face smelling upwards multiple an colorful florals roses hats above many paints with earrings turmeric makeup brightly orange red pink wth scattered among yellow oranges flying flying butterflies teal background on teal blue background lips eyebrow hadid cg poster
Application #1

Prompt for Synthetic Images

an illustration albu etching vsccam illustration intricate insect heavily black intricate intricate insect insect crest intricate crest on an behind lit circular moon intricate folk os intricate insect insect form a exhibiting called an intricate insect shown frontal frontal surrounded amongst an lit many crescent moons besides scattered stars and stars and moons past gold beige navy amongst beside among and crescent beside and crescent navy stars on dark navy background night stars bohemian etching logo
Visual Question Answering

[Prompt] Answer the question given the context.

Question: What other big vehicle is often painted about the same shade as this vehicle?
Visual Question Answering

[Prompt] Answer the question given the context.

A colcandidenver lantic closeup former recent train train parked tradition enclosed metrotram in a red livery it on railroad platform containing wearing a yellow pol surround a knob beside platform near a under platform shelter right there and roof shadows and platform and tracks etc wore worn worn mau maroon brown white stripes markings content worn yellow yellow stripes train pretoria namibia railway platform train operator worn brown windows platform platform

Question: What other big vehicle is often painted about the same shade as this vehicle?
Visual Question Answering with LLM

[Prompt] Answer the question given the context.

a colcandidenver lantic closeup former recent train train parked tradition enclosed metrotram in a red livery it on railroad platform containing wearing a yellow pol surround a knob beside platform near a under platform shelter right there and roof shadows and platform and tracks etc wore worn worn mau maroon brown white stripes markings content worn yellow yellow stripes train pretoria namibia railway platform train operator worn brown windows platform platform

Question: What other big vehicle is often painted about the same shade as this vehicle?

# Multi-Modal Few-Shot Learning

<table>
<thead>
<tr>
<th>Input Prompt</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /> Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.</td>
<td><img src="image2.png" alt="Image" /> A portrait of Salvador Dali with a robot head.</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /> Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone.</td>
<td><img src="image4.png" alt="Image" /> Je suis un cœur qui bat pour vous.</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /> pandas: 3</td>
<td><img src="image6.png" alt="Image" /> giraffes: 4</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image" /> I like reading, my favourite play is Hamlet. I also like</td>
<td><img src="image8.png" alt="Image" /> Dreams from my Father.</td>
</tr>
</tbody>
</table>

Flamingo, DeepMind, 2022
**Application #2**

## Modular, Better

<table>
<thead>
<tr>
<th>Model</th>
<th>VQAv2</th>
<th>OKVQA</th>
<th>COCO Caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flamingo-80B</td>
<td>56.3/67.6</td>
<td>50.6/57.8</td>
<td>84.3/113.8</td>
</tr>
<tr>
<td>De-Diffusion with PaLM 2</td>
<td>67.2/68.4</td>
<td>57.0/60.6</td>
<td>88.5/103.7</td>
</tr>
</tbody>
</table>

Representing images with text provides *flexibility* for using LLMs for multimodal tasks.

- **easy exploitation** of advanced LLMs: no joint training required, just inference;
- **easy data collection**: no interleaved image-text data required, just images.
Application #3

Interactions with ChatGPT

Given an image with keywords:
“an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street”

What is unusual about this image?
It is unusual to see a corgi riding a bicycle, especially in a busy city like New York…”
More examples

Photos

a colcord vindic analogue closeup of yellow bus mirror through a mirror mirror mirror depicting a yellow vehicle it on black bicycle foreground behind a black pol handle a boards behind blur towards a white boards boards above background silhousky sky white siding and poles pole besides wooden pole lb orange green green handle reflection content reflection black black round mirror rotterdam wisconsin congestion vehicle mirror mirror mirror vintage hdr reflection glass mirror
More examples

Creative Arts

an hdr painting robotic cat aus blackandwhite steampunk cyborg cat with eye with crook standing an on black leash robot steampunk steampunk os steampunk punk robot resembling called an filly cat shown standing neck showcasing relating an derel steampunk mechanism robot with optom eye optomeyes optomelenses silver monochrome brown monochrome consist overlooking with derel mechanism overlooking scratched cities faded codes on beige yellow background portrait eyebrow steampunk steampunk artwork
More examples

Abstract Symbols

a coleman vsocam appreciation an lions animal walking side winding winding lion on a beige pared it minimalist profile silhouette featuring featuring a black pol curled a silhouette atop lineup in a atmospheric silhouette silhouette on pale beige pared backgrounds grey background minimalist minimalist minimalist gladly featuring minimalist silhouette black and blk navy black white symbol modern i modernist minimalist white lions lions atx wsj fintech minimalist line symbol render ometric silhouette profile symbol
More examples

**Factual Knowledge**

a colalbugmbearts drawing beyonce beyonce portrait woman portrait with swept swept curls in a silver dress it with colourful watercolor there with a a colorful pollouda cloud above atop with a colorful watercolor graffiti atop beige beige background shadows grey background with spots wth with woman ear bold magenta purple black sleek off sleeveless black and white black and white yellow swoo curls rihanna jimi supermodel abstract inktober drawing illustration stration face eyebrow portrait
More examples

**Factual Knowledge**

- A colorful bugmearts drawing Beyonce Beyonce portrait woman portrait with swept swept curls in a silver dress it with colourful watercolor there with a colorful polouda cloud above atop with a colorful watercolor graffiti atop beige background shadows grey background with spots with woman ear bold magenta purple black sleek off sleeveless black and white black and white yellow swoo curls Rihanna Jimi supermodel abstract inktober drawing illustration stration face eyebrow portrait
Factual Knowledge

A typical Grand Canyon, many brown vast peaks mountains and canyon and overview surrounded an overview mostly overview canyon grand grand grand grand canyon shown foreground note an overview canyon shown overview lineup whilst towards an morning morning dusk overview also morning morning blue sky and sky redness orange brown purple with numerous towards vast cliffs numerous Grand Canyon grand cliffs mountains mostly blue background dusk overview grand grand Canyon.
More examples

**Factual Knowledge**

A cartoon illustration of Cinderella from Disney's animation. Cinderella is wearing a pale dress with sparkling webs and a silver polka-dotted string. She is surrounded by darkness and purple background with shadows. She has yellow hair in a ponytail and is featured in a hime gown with a white face.