Retrieval-augmented Multimodal Foundation Models

Michihiro Yasunaga
Stanford University
AI is becoming multimodal

Personal Assistants

Search

Generative AI

Autopilot
Multimodal Foundation Models (Text-to-Image)

DALL·E, Parti (text → image; Transformer)
DALL·E 2, StableDiffusion (text → image; Diffusion)

A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!
### Multimodal Foundation Models (Image-to-Text)

**Flamingo, GPT-4** (image → text; Transformer)

<table>
<thead>
<tr>
<th>Input Prompt</th>
</tr>
</thead>
</table>
| ![Chinchilla](image1)  
This is a chinchilla. They are mainly found in Chile. |
| ![Shiba](image2)  
This is a shiba. They are very popular in Japan. |
| ![Flamingo](image3)  
This is a flamingo. They are found in the Caribbean and South America. |
| ![Propaganda Poster](image4)  
Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese. |
| ![Pink Room](image5)  
Output: A pink room with a flamingo pool float. |
| ![Portrait](image6)  
Output: A portrait of Salvador Dali with a robot head. |
| ![Pandas](image7)  
pandas: 3 |
| ![Dogs](image8)  
dogs: 2 |
| ![Giraffes](image9)  
giraffes: 4 |
Multimodal Foundation Models (Unify Text & Image)

CM3 (text ⇝ image; Transformer)

Unified VLM

Text & image generation

Labrador sitting near water.

Labrador sitting near water.
However, models may lack knowledge and **hallucinate**.
Challenge

Current models’ knowledge is bounded by the parameters & training data. Can we allow models to refer to external memory?

- What does an Armenian church look like?
- What is the name of this place?
- The Dragon and Tiger Pagodas next to fireworks.
- Empire state building and fireworks
Who is the president of the US?

Generator
(Language Model)

Joe Biden

Retriever

Memory

WIKIPEDIA
The Free Encyclopedia

Joe Biden is the 46th and current president of the United States, assumed office on January 20, 2021.

Retrieved document

Inspiration: Retrieval-augmented Language Model

Guu+2020; Lewis+2020
Who is the president of the US?

Retriever

Can expand & update knowledge (e.g. new domain, news)

Memory

Retrieved document

Generator (Language Model)

Joe Biden

More accurate

Joe Biden is the 46th and current president of the United States, assumed office on January 20, 2021.
Retrieval-augmented multimodal modeling

RA-CM3: Retrieval-augmented multimodal modeling.
Our Idea: Retrieval-augmented Multimodal Model

Retriever

Labrador sitting on bench near water.

Generator (Multimodal Model)

Labrador retriever sitting on bench.

Multimodal memory

Multimodal “document”: image, text, or mixture of them
Our Idea: Retrieval-augmented Multimodal Model

Text-to-Image

Retriever

Labrador sitting on bench near water.

Generator (Multimodal Model)

Labrador retriever sitting on bench.
Our Idea: Retrieval-augmented Multimodal Model

Image-to-Text

Multimodal memory

Retriever

Generator (Multimodal Model)

Labrador sitting on bench near water.

Labrador retriever sitting by water.
Technical innovations

- What is an effective **multimodal retrieval** method?
- How to **integrate** retrieved items into the **generator**?

![Diagram showing the process of multimodal retrieval and integration](image-url)
Multimodal Retrieval

Labrador sitting on bench near water.

Generator
(Multimodal Model)

Retriever

Multimodal memory

Labrador retriever sitting on bench.
Multimodal Retriever

Dense Retriever with Mix-modal Encoder

\[ f(\text{query}, \text{memory}) \rightarrow \text{score} \]
Background: CLIP

CLIP produces text embeddings and image embeddings in shared vector space
Multimodal Retriever

Dense Retriever with Mix-modal Encoder

\[ f(\text{query, memory}) \rightarrow \text{score} \]
Example

Query

Labrador sitting on bench near water.

Labrador retriever sitting on bench. 0.85
Labrador retriever sitting by water. 0.81

Labrador retriever sitting on bench. 0.35
Strategy for Retrieval

Relevance
The retrieved items should be relevant to query

✅ Cosine similarity score + Maximum Inner Product Search

Diversity (for training)
If simply take items of top scores, may include duplicate images/text
This can cause the generator to overfit or learn repetitive generation

💡 Avoid redundant items
  ■ Skip candidate item if it is too similar to query or items already retrieved

💡 Query dropout
  ■ Drop some tokens of query used in retrieval (e.g. 20% of tokens)
  ■ This further increases diversity and serves as regularization

Diversity is crucial in multimodal setting
- Multimodal dataset often contains duplicate images across docs
- Each image takes many tokens (1024), so can significantly hurt model training

Can improve FID score by 5 points
Labrador sitting on bench near water.

Retriever

Generator (Multimodal Model)

Labrador retriever sitting on bench.
Generator: Retrieval-Augmented CM3

Causal masked language model (CM3)

Transformer

Retrieved item 1
Retrieved item 2
Main input

Labrador retriever sitting on bench.
Labrador retriever sitting by water.
Labrador sitting on bench near water.

Each image is tokenized into 1024 tokens using VQ-VAE
Train the Generator Efficiently

$$\text{Loss} = (\text{LM loss for main input}) + \alpha \cdot (\text{LM loss for retrieved items})$$

- Existing retrieval augmented LMs: $\alpha = 0$
- Our method: $\alpha > 0$ ($\alpha = 0.1$ works the best)

$\alpha > 0$ has effect like increasing batch size without extra forward compute, increasing training efficiency.

$\alpha > 0$ is crucial in multimodal setting
- Each image takes many tokens (1024)
- If $\alpha = 0$, we are throwing away a lot of compute

Can reduce training time by 50%
Retrieval Augmented Multimodal Model

Labrador sitting on bench near water.

Labrador retriever sitting on bench.
## Comparison with related models

<table>
<thead>
<tr>
<th>Model</th>
<th>Image Generation</th>
<th>Text Generation</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>DALL-E, StableDiffusion, Imagen, etc.</td>
<td>✅</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>kNN-diffusion, Re-Imagen, etc.</td>
<td>✅</td>
<td>-</td>
<td>✅</td>
</tr>
<tr>
<td>Flamingo, GPT-4, etc.</td>
<td>-</td>
<td>✅</td>
<td>-</td>
</tr>
<tr>
<td>MuRAG, Re-ViLM, REVEAL, SmallCap, etc.</td>
<td>-</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>CM3</td>
<td>✅</td>
<td>✅</td>
<td>-</td>
</tr>
<tr>
<td>RA-CM3 (Ours)</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
</tbody>
</table>
Experiments

Train data
- **LAION** (cleaned 150M image-text pairs)
  External memory: LAION

Evaluation
- **MSCOCO** caption2image, image2caption.
  External memory: MSCOCO train set

Model
- Transformer with seq_length 4096 (up to 2 retrieved documents)
- 2.7B parameters trained for 5 days on 256 GPUs
- “**Retrieval Augmented CM3 (RA-CM3)**”

Baseline
- Vanilla CM3 with no retrieval, same size, trained using the same amount of compute
### Performance (Text-to-Image)

Retrieval improves caption-to-image generation quality (e.g. RA-CM3 vs CM3)

<table>
<thead>
<tr>
<th>Model</th>
<th>Model type</th>
<th>#Train images</th>
<th>MSCOCO FID score (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DALL-E (12B)</td>
<td>Autoregressive</td>
<td>250M</td>
<td>28</td>
</tr>
<tr>
<td>Parti (20B)</td>
<td>Autoregressive</td>
<td>6B</td>
<td>7.2</td>
</tr>
<tr>
<td>Stable Diffusion</td>
<td>Diffusion</td>
<td>1B</td>
<td>~12</td>
</tr>
<tr>
<td>Vanilla CM3</td>
<td>Autoregressive</td>
<td>150M</td>
<td>29</td>
</tr>
<tr>
<td>RA-CM3</td>
<td>Autoregressive</td>
<td>150M</td>
<td>16</td>
</tr>
</tbody>
</table>

An Armenian church.

13 points improvement
## Performance (Image-to-Text)

Retrieval improves image-to-caption generation quality (e.g. RA-CM3 vs CM3)

<table>
<thead>
<tr>
<th>Model</th>
<th>#Train images</th>
<th>MSCOCO CIDEr score (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parti (20B)</td>
<td>6B</td>
<td>0.84</td>
</tr>
<tr>
<td>Flamingo (3B) 4-shot</td>
<td>2.5B</td>
<td>0.85</td>
</tr>
<tr>
<td>Vanilla CM3</td>
<td>150M</td>
<td>0.72</td>
</tr>
<tr>
<td>RA-CM3</td>
<td>150M</td>
<td>0.89</td>
</tr>
</tbody>
</table>

17 points improvement

Image to text

The Dragon and Tiger Pagodas next to fireworks.
Retrieval improves training efficiency

- RA-CM3 outperforms DALL-E while using only 30% of training compute
Accurate Image Generation

**RA-CM3 Retrieved items**
- French flag
- Oriental Pearl tower

**RA-CM3 outputs**
- Input: “French flag waving on the moon’s surface.”
- Input: “The Oriental Pearl tower in oil painting.”

**Baseline outputs**
- (Vanilla CM3)
- (Stable Diffusion)
Accurate Image Generation

**Input:** “An Armenian church during a sunny day.”

**RA-CM3 outputs**

* Armenian church

**Baseline outputs**

*(Vanilla CM3) (Stable Diffusion)*

**Input:** “Photo of the Statue of Liberty standing next to the Washington monument.”

**RA-CM3 Retrieved items**

* Statue of Liberty
* Washington monument

**Baseline outputs**

*(Vanilla CM3) (Stable Diffusion)*
Accurate Image Generation

RA-CM3 Retrieved items

Ming Dynasty vase

RA-CM3 outputs

Input: “A Ming Dynasty vase with orange flowers painted.”

Baseline outputs (Vanilla CM3) (Stable Diffusion)

Ming Dynasty vase

RA-CM3 Retrieved items

Mount Rushmore

Japanese cherry

Accurate Image Generation

**RA-CM3 Retrieved items**

**RA-CM3 outputs**

- Input: “Photo of the Callanish standing stones, fireworks in the sky.”
- Input: “Photo of the Dragon and Tiger Pagodas, the sun is setting behind.”

**Baseline outputs**

(Vanilla CM3)  (Stable Diffusion)
Multimodal In-Context Learning

RA-CM3
In-context

(Demonstrate the style to generate)

RA-CM3 output

“Photo of a house taken on an autumn day.”

Baseline outputs

(Vanilla CM3)

(Stable Diffusion)

Intuition:
After retrieval augmented training, our generator model has learned how to use in-context examples and acquired this in-context learning capability.

(Demonstrate the style to generate)

“Painting of red roses.”
Image Editing

Provide an image to control the type of editing

Source image  Masked image  RA-CM3 output

RA-CM3 In-context
Image Editing

Source image | Masked image | RA-CM3 In-context | RA-CM3 output

RA-CM3

RA-CM3

RA-CM3
One-shot Image-to-Text

Task

animal X  animal Y  animal __

Result

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline CM3</td>
<td>0.53</td>
</tr>
<tr>
<td>RA-CM3</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Motivation: test the true in-context learning capability of our generator
**Few-shot Image-to-Text**

**Ensemble (e.g. 2)**

Result

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Ensembles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Baseline CM3</td>
<td>0.53</td>
</tr>
<tr>
<td>RA-CM3</td>
<td>0.78</td>
</tr>
</tbody>
</table>

**Takeaway:**
- Generator exhibits good in-context learning performance
- Ensemble is an effective method to increase in-context examples
Summary

RA-CM3: The first retrieval-augmented multimodal model that can retrieve and generate both text and images

Result & Impact: Retrieval enables

- Accurate image/text generation ⇒ reduce hallucination
- Efficient training ⇒ reduce cost of training large foundation models
- Multimodal in-context learning (e.g., can prompt using both images and text)
Thank you!

https://cs.stanford.edu/~myasu/

@michiyasunaga