Multimodal RAG

gollnickdata.de

Multimodal RAG

Introduction

What is it?

- Retrieval based on different content types
- Typically text and image

Why is it important?

- Only text might not be enough for some analysis
- Helps to understand complex concepts with visual support

Multimodal RAG

Approaches

Joint Embeddings

Embedding model suitable for text and images simultaneously

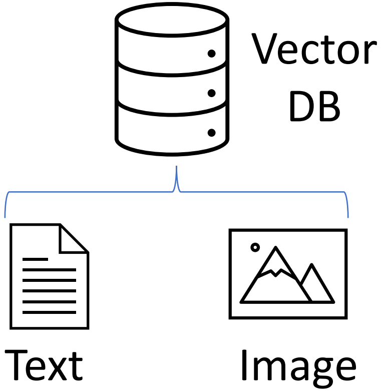
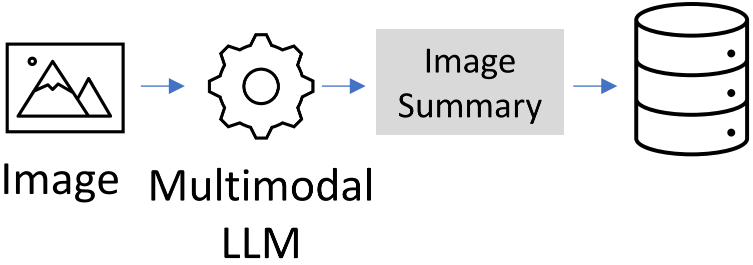


Image-to-Text

Multimodal models generate summaries from images



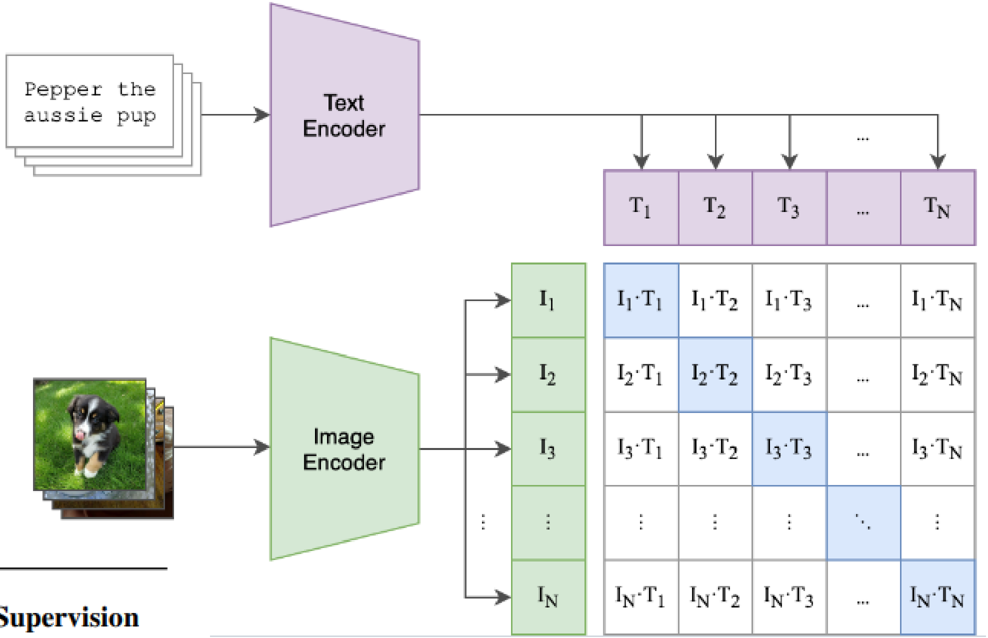
gollnickdata.de

Multimodal RAG

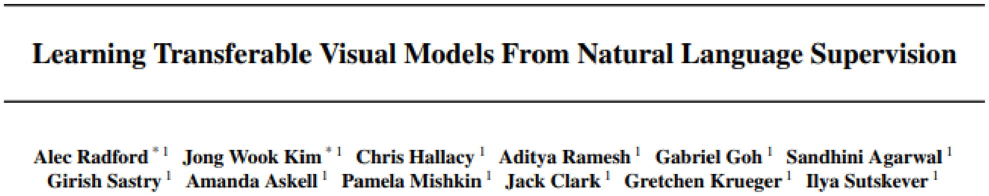
Joint Embeddings

Embedding models

- CLIP (Contrastive Language-Image Pre-Training



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$



Learning Transferable Visual Models From Natural Language Supervision

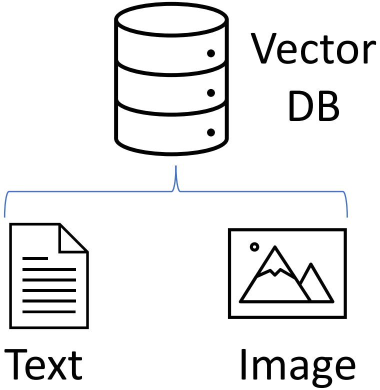
Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Source: https://arxiv.org/pdf/2103.00020

Multimodal RAG

Coding

Data Ingestion



Vector DB

Text

Image

Flickr8k Dataset

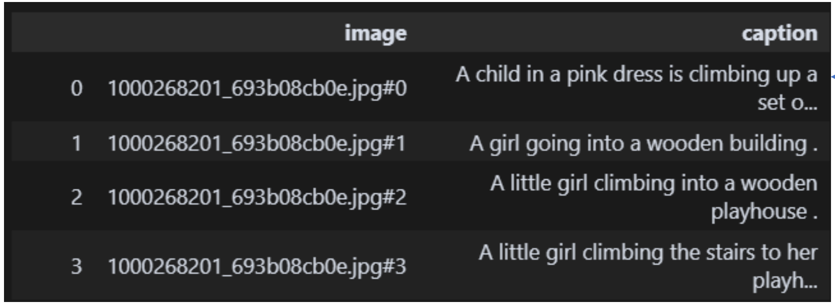
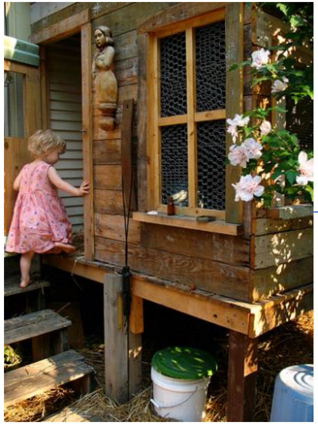


	image	caption
0	1000268201_693b08cb0e.jpg#0	A child in a pink dress is climbing up a set o...
1	1000268201_693b08cb0e.jpg#1	A girl going into a wooden building .
2	1000268201_693b08cb0e.jpg#2	A little girl climbing into a wooden playhouse .
3	1000268201_693b08cb0e.jpg#3	A little girl climbing the stairs to her playh...



gollnickdata.de