Better Retrieval for Generation

CS6803 - Topics in NLP
Dr. Maunendra Sankar Desarkar

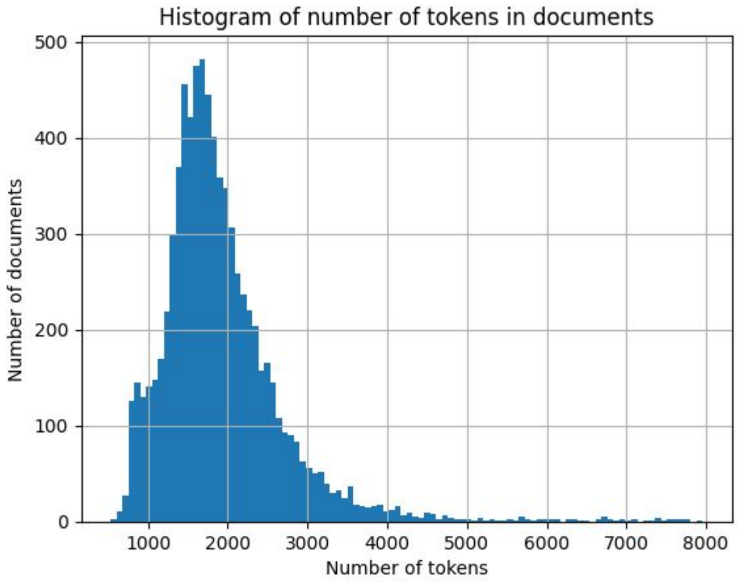
By:
Aaryan - CO21BTECH11001
Abhishek Kumar - AI21BTECH11003

Work Allocation and Index

- Aaryan: **Fine-tuning and evaluation of retrievers on domain-specific data**
  - Dataset and EDA
  - Evaluation Setup
  - Evaluation of pre-trained models
  - Fine-tuning strategies
  - Current Results
- Abhishek: **Multimodal RAG**
  - Literature review (FROMAGe paper and uniMUR paper)
  - Experimentation with FROMAGe to identify its limitations

Dataset and EDA

- MedQuAD dataset:
  - **7873** reachable document urls
  - **36925** questions
- Scraping and cleaning of webpages
  - Using **BeautifulSoup**
  - Removed extra line characters and whitespaces
- Number of tokens in documents (nomic tokenizer):
  - Min: **531**, Max: **7952**
  - Mean: **1918**, Median: **1770**



MedQuAD: Ben Abacha, A., Demner-Fushman, D. A question-entailment approach to question answering. BMC Bioinformatics 20, 511 (2019). https://doi.org/10.1186/s12859-019-3119-4

Evaluation Setup

- If required, documents are **chunked**
- Embeddings of documents are **pre-computed** and stored
- Chunks are retrieved using **cosine-similarity**
- Metrics
  - Recall @ k

$$\text{Recall}-k = \frac{\text{Number of relevant chunks in top } k}{\text{Number of relevant chunks}}$$

  - MRR @ k

$$\text{MRR}-k = \sum_{\text{chunk} \in \{\text{relevant chunks}\}} \frac{1}{\text{Index of chunk in top k retrieved chunks}}$$

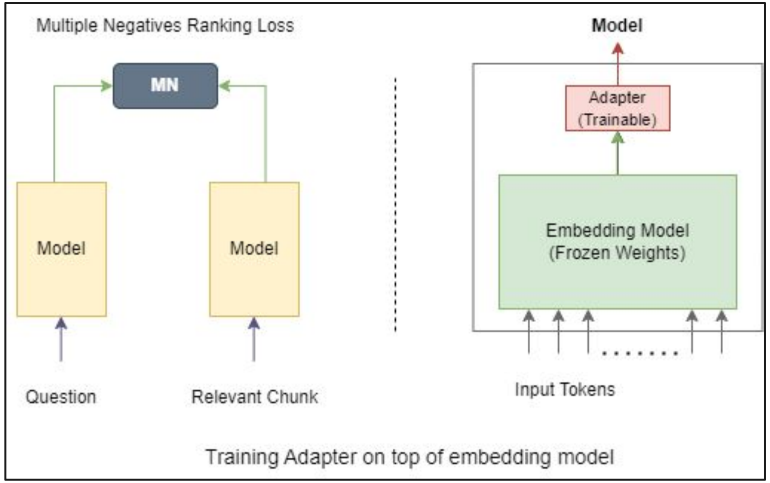
Evaluation of pre-trained models

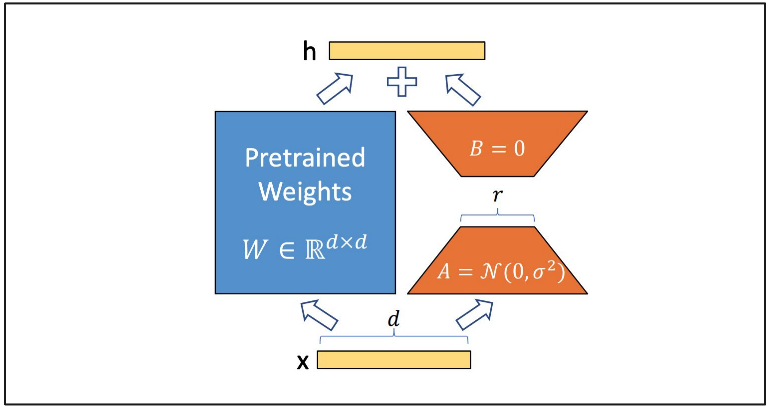
- paraphrase-mpnet-base:
  - Context Window: 512, 109M parameters
  - 33545 chunks formed
- nomic-embed-text , bge-m3
  - Context Window: 8192
  - No need of chunking

Model Name	Recall				MRR			
	Recall@1	Recall@3	Recall@10	Recall@100	MRR@1	MRR@3	MRR@10	MRR@100
paraphrase-mpnet-base	19.55	38.02	52.22	66.98	78.83	84.79	85.57	85.62
nomic-embed-text	88.6	97.76	**99.19**	**99.69**	89.02	92.98	93.18	93.18
bge-m3	**89.58**	**97.99**	99	99.56	**99.02**	**93.64**	**93.78**	**93.79**

Fine-tuning strategies

- Adapter fine-tuning
  - Freeze the model weights
  - Train a light-weight adapter on top
  - Multiple Negatives Ranking Loss used
  - 8 mins per epoch (1.4 GB)
- LoRA
  - Freeze the model weights
  - Add matrices of low rank
  - Cosine Similarity Loss used
  - 14 mins per epoch (20 GB)





LoRA Diagram Source: https://heidloff.net/article/efficient-fine-tuning-lora/

Current Results

- Used **20k** queries for training and **5k** for testing
- Two layer Adapter
  - Small improvements
- LoRA
  - Poor performance than pre-trained
  - Possible reason: Catastrophic Forgetting

Model Name		Recall				MRR			
		Recall@1	Recall@3	Recall@10	Recall@100	MRR@1	MRR@3	MRR@10	MRR@100
Adapter	Training	19.67%	38.17%	54.03%	69.20%	79.50%	85.12%	85.90%	85.93%
	Testing	20.18%	36.86%	50.54%	65.19%	78.04%	84.34%	85.31%	85.43%
LoRA	Training	0.46%	0.96%	2.11%	8.90%	1.82%	2.65%	3.39%	4.11%
	Testing	0.47%	1.04%	2.27%	8.03%	1.74%	2.67%	3.38%	3.93%

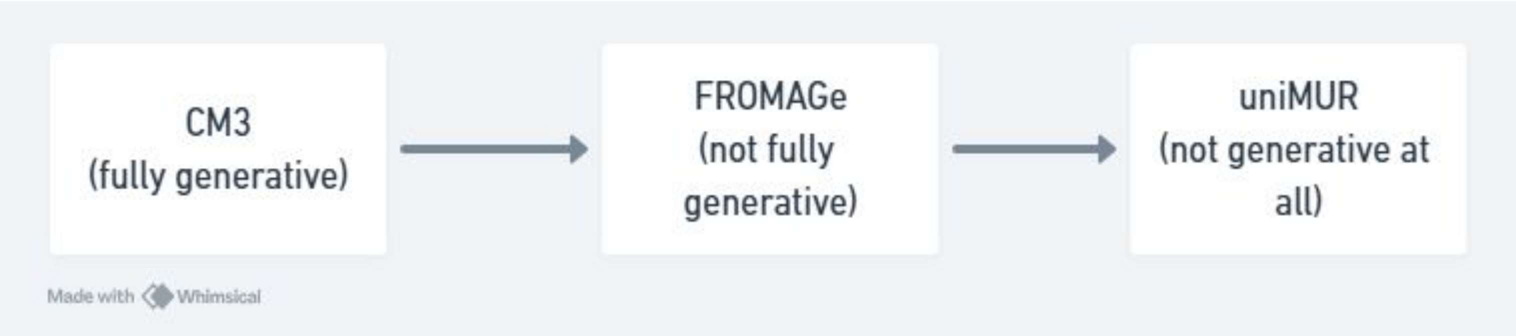
Upcoming experiments

- Finetune bigger adapters to check generalization
- Try different loss functions while fine-tuning using LoRA
- Try other approaches like QLoRA / OLoRA

Literature Review and a bit of experimentation

Sources :-
https://arxiv.org/pdf/2301.13823 (fromage)
https://aclanthology.org/2024.findings-eacl.105/ (uniMUR)

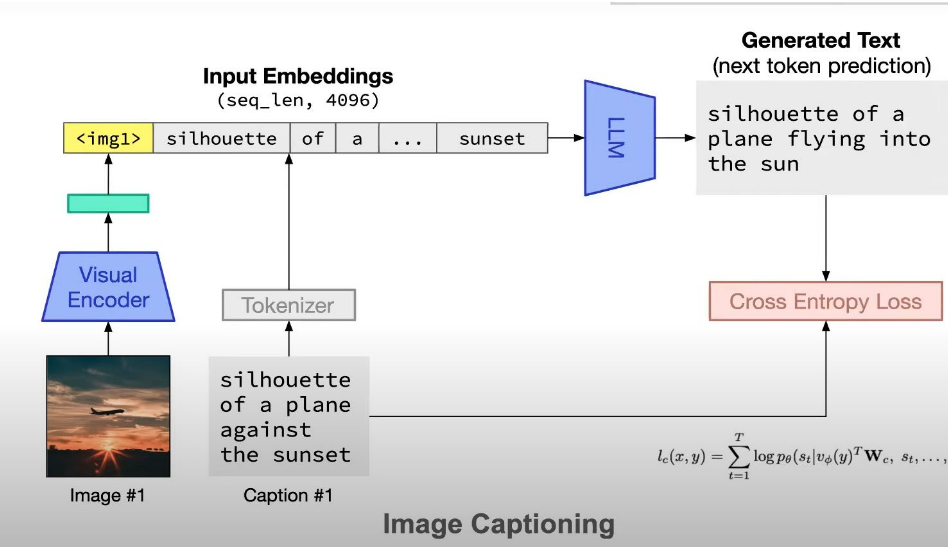
Multimodal RAG (I/O - image and text interleaved)

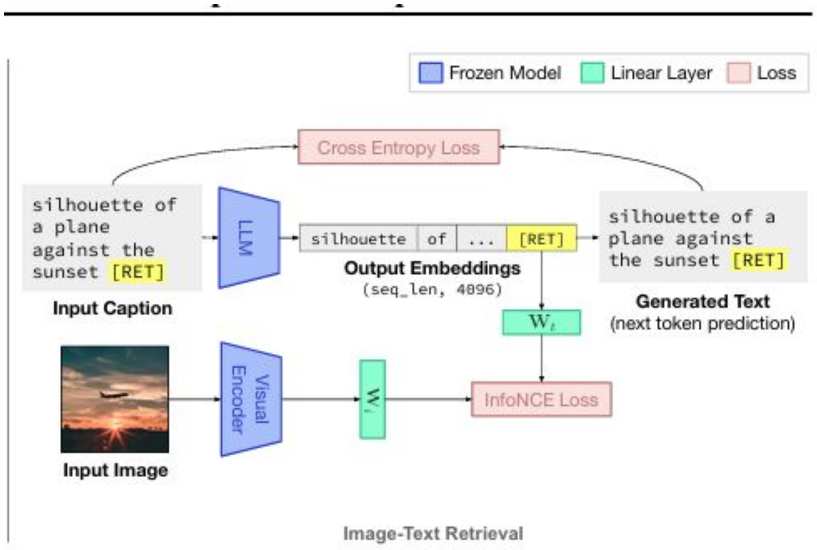


Casual Masked Multimodal Model of the Internet

1. Earliest prior work - proposing Model with Multimodal I/Os
2. Generally not available to the public
3. With 384 GPUs(nvidia A100 model) , training for 24 days - large computational resources
4. Poor performance on VIST (visual storytelling text-image) dataset .Most outputs produced by CM3 are not interpretable or relevant wrt to their inputs .

Frozen Retrieval Over Multimodal Data for Autoregressive Generation





FROMAGe

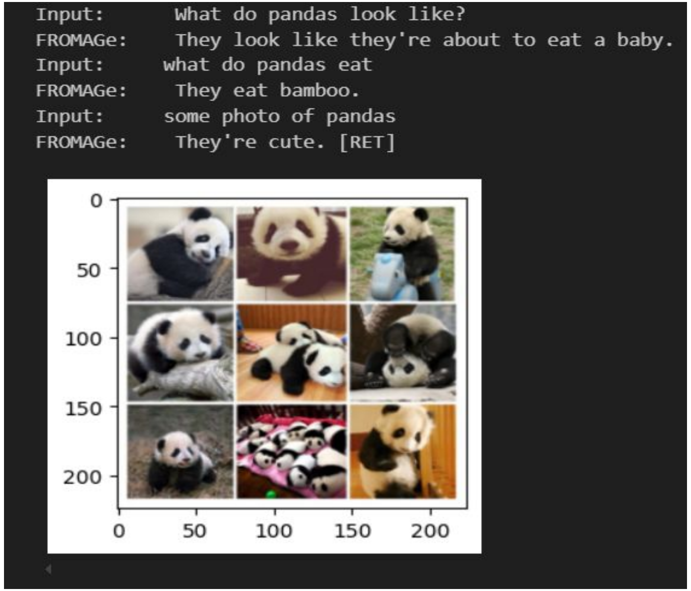
1. Open source
2. 97% parameters frozen (leverages pre-trained LLM) . hence computationally more efficient .Single gpu(nvidia A100) , training for 1 day .
3. Generates outputs semantically meaningful wrt inputs . thus ,outperforms CM3.
4. Language modelling and contrastive learning objectives.
5. It can handle a variety of zero shot and few shot tasks.

Limitations :-

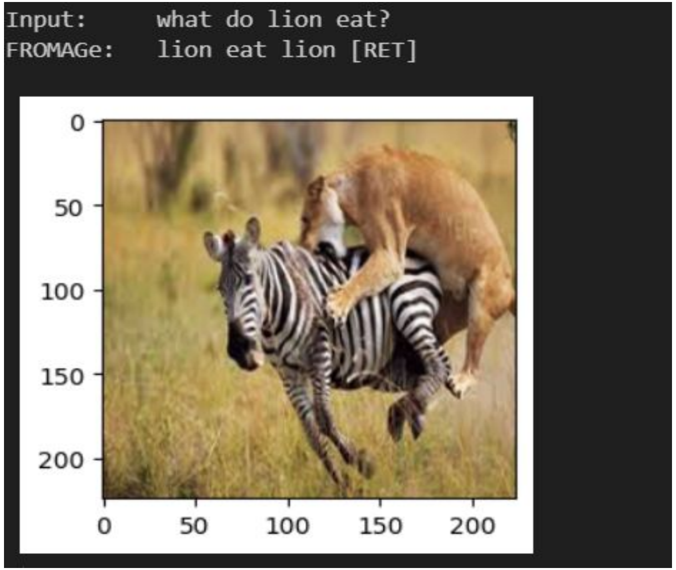
Fromage exhibits a stronger bias towards generating text only tokens - avoiding [RET] token (primarily used to retrieve the relevant image) because of LLM bias not to generate [RET] token (generating text only outputs)

Experimentation with FROMAGe - limitations

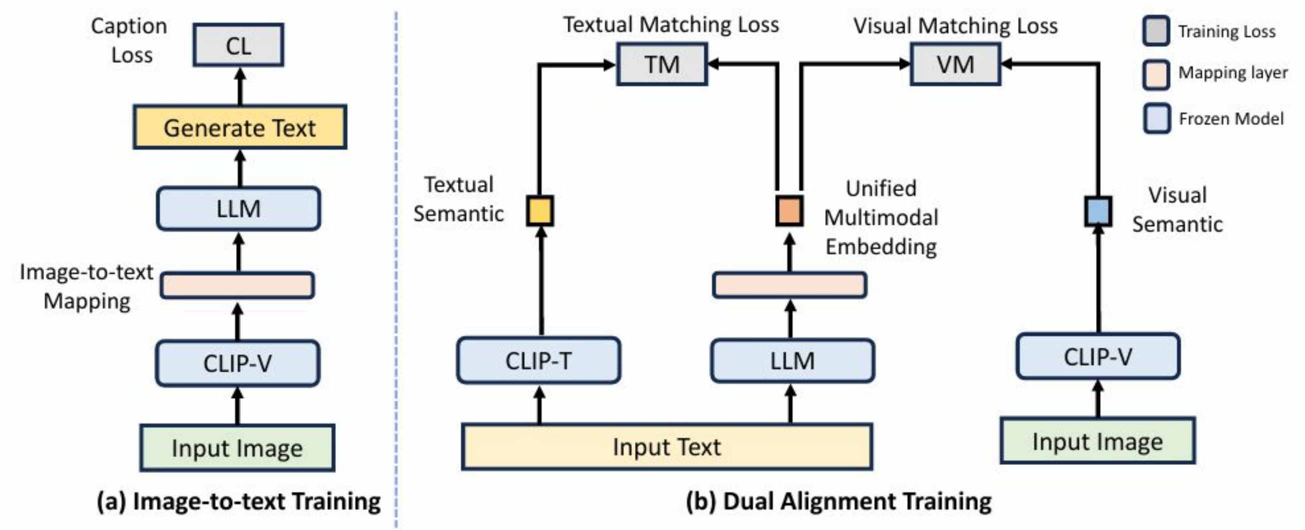
a)Text bias of the LLM



b) Inconsistent -image with text outputs



Unified Embeddings for Multimodal Retrieval



uniMUR

1. closed source
2. 98% frozen parameters , thus even more computationally efficient than FROMAGe.4 * V100 GPU , training for less than 16 hours .
3. Mitigates the text only bias of the FROMAGe using unified embeddings.Also the coherency increase between image output and text output and overall outputs and the input. thus outperforms FROMAGe.
4. Language modelling and contrastive learning objectives
5. It is not optimised for few shot tasks directly . focus on zero shot tasks .

Future Work

- Implementing uniMUR’s logic
- Training ,testing and verifying the details mentioned in the paper like recall etc
- Finally making the uniMUR’s code open source