
Multimodal Document Question Answering on Scholarly Articles

Anirudh Joshi

Information Networking Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ajoshi3@andrew.cmu.edu

Manigandan Ramadasan

Information Networking Institute
Carnegie Mellon University
Pittsburgh, PA 15213
mramadas@andrew.cmu.edu

Yiwei Chen

Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
yiweiche@andrew.cmu.edu

Abstract

Text-based Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of natural language tasks; however, they often struggle when processing multimodal documents—such as research papers and financial reports—that incorporate information in images and tables. In this work, we aim to enhance the performance of smaller LLMs for multimodal document Question Answering (QA), with a particular focus on research papers. We explore Image-based Retrieval-Augmented Generation (RAG) techniques as in a real-world QA system utilizing an open-source model with a low context length there is no possibility to pass the entire paper, instead the system must pick the most relevant image based on the user query. Our experiments, design, and code are available in our GitHub repository¹.

1 Introduction

1.1 Motivation

The rapid advancements of LLMs have led to a significant improvement in question-answering systems, yet the widespread adoption of LLMs for question-answering systems has been limited by computational demands. When documents are fetched from online sources such as research papers, it presents a limitation in terms of the amount of information captured as we miss out on key information present in other modalities such as images and tables when we use LLMs that depends on text. Another reason is that large LLMs is expensive due to their computational resources usage. Thus, there is a need for smaller LLMs that can perform QA on documents in a multimodal fashion. To address this, we will incorporate Image-based RAG to improve the retrieval capabilities of smaller models, allowing them to extract relevant information from figures and tables within research papers.

To our knowledge currently there hasn't been significant work done in the domain of improving Image RAG models for Multimodal QA on research papers.

¹<https://github.com/joshianirudh/IDL-Project>

1.2 Objective

The objective of this project is to improve the performance of smaller LLMs on document Q&A. Specifically, we will focus on Q&A for research papers making sure that the models can handle multimodal content. Given the difficulty in improving performance across different modalities, we will be using Image-based RAG to enhance the models' ability to retrieve and reason over figures and tables which leads to more accurate and context aware responses.

The current **SPIQA dataset has a major drawback**, it provides the model with the relevant image from the dataset to answer the question, but in a real-world system the model will not have this information.

2 Literature Review

SPIQA [1] is the foundation of our project work as we will be using it for benchmarking and fine-tuning our models. This paper had several key results which will guide our research such as image captions improve the performance, open models tend to struggle with reasoning on tables and lowering image resolution to accomodate the low context-lengths of open-models leads to significant performance degradation. We plan on working on these aspects to enhance the performance of small open-source models.

Recent work shows that fine-tuning multimodal transformers can degrade cross-modal alignment [2], especially with limited data. Full fine-tuning can lead to overfitting and poor generalization. Augmenting a frozen or lightly adapted model with retrieval improves performance. Their experiments validate that retrieval-augmented generation (RAG) maintains better modality alignment in multimodal QA tasks.

The RAG survey paper by Gao et al [3] highlights the effectiveness of RAG approaches in reducing hallucinations and their implication on applications requiring factual corectness.

Contrastive Language Image Pre-training (CLIP) [13] uses contrastive learning to learn the relationship between an image and a text where they are brought closer in a shared embedding space while pushing apart mismatched pairs during training. CLIP consists of image and text encoders and it is capable of performing zero shot image classification by comparing an image to text descriptions of possible labels without needing any further fine tuning.

ColBERTv2 [14] is an information retrieval model that uses late interaction by encoding queries and documents where each token in a sentence is embedded into embedding using BERT and it uses MaxSim to compute the maximum similarity between each query token and all passage tokens and then it sums these to score relevance.

3 Methodology

3.1 Model Description

Current research indicates that fine-tuning for a specific task can lead to a loss of reasoning ability in large language models (LLMs) and may result in worse overall performance. Consequently, the current literature often favors Retrieval-Augmented Generation (RAG) over full fine-tuning.

For this project we will focus on Image RAG as current Text RAG models are already capable in their abilities.

1. **Image-RAG:** We propose to train a CLIP model to retrieve images based on a question (user query) for this we will first create a RAG dataset for contrastive training, with a question, positive image anchor and a negative image anchor. The CLIP model parameter count, input and output shape can be seen in Figure 1.

While training CLIP, we created the RAG dataset triplets using the SPIQA train set. The SPIQA train set questions are designed in such a way that the question can only be answered using the reference_figure provided in the dataset, out of all the figures in the research paper. We initialize the positive image anchor as the reference_figure and randomly sample the negative image anchor

Layer (type:depth-idx)	Param #
└─CLIPTextTransformer: 1-1	--
└─CLIPTextEmbeddings: 2-1	--
└─Embedding: 3-1	25,296,896
└─Embedding: 3-2	39,424
└─CLIPEncoder: 2-2	--
└─ModuleList: 3-3	37,828,608
└─LayerNorm: 2-3	1,024
└─CLIPVisionTransformer: 1-2	--
└─CLIPVisionEmbeddings: 2-4	--
└─Conv2d: 3-4	2,359,296
└─Embedding: 3-5	38,400
└─LayerNorm: 2-5	1,536
└─CLIPEncoder: 2-6	--
└─ModuleList: 3-6	85,054,464
└─LayerNorm: 2-7	1,536
└─Linear: 1-3	393,216
└─Linear: 1-4	262,144
Total params: 151,276,544	
Trainable params: 151,276,544	
Non-trainable params: 0	

Figure 1: CLIP Model Summary

from the list of all figures in the paper excluding the reference figure. We use Triplet Margin Loss to fine-tune CLIP.

Other than CLIP, we also use pre-trained ColBERTv2 for text-based retrieval. We chose ColBERTv2 for text-based retrieval because it uses token level embeddings rather than single embedding for the whole sentence which leads to loss of information. On top of that, it also uses late interaction between query and document embedding which increases the efficiency when dealing with long documents. We integrate ColBERTv2 through the LangChain framework.

3.2 Dataset

In this project, we are using Scientific Paper Image Question Answering (SPIQA) dataset [1] by Google, which contains 270,194 QA samples that have 152,487 figures, 117,707 tables, captions and text. The samples in the dataset are derived from 25,859 papers from 19 Computer Science (CS) conferences (this also means that we are narrowing our scope to CS research papers). To generate QAs in SPIQA, the authors used Gemini 1.5 Pro. The dataset is divided into training, validation, and three test sets: test-A, test-B and test-C. test-B and test-C consist of human written QAs that are derived from the QASA and QASPER datasets. The reason we chose this dataset is because it is a relatively new data set and it has a diverse range of questions with variations in length and complexity that require strong reasoning capability.

For evaluating Gemini 1.5 Pro and InstructBLIP, we used test-A which is the largest test set that has 666 QAs from 118 research papers. We chose test-A over test-B and test-C because the QAs from test-A are taken from every conference making it diverse set and it also means that the multimodal questions covers a broader range of figures, tables, making it a balanced test set for evaluating the baseline models.

3.3 Evaluation Metric

We use top-k retrieval accuracy for evaluating our CLIP model. This metric measures the proportion of queries for which the correct item is found within the top k retrieved results. A higher top-k accuracy means that the CLIP is able to rank relevant results near the top of the retrieval list. We use top-k retrieval accuracy because it tells how often the correct document appears in top k results and we use it to determine if CLIP is able to retrieve the right document close to the top.

$$\text{top - k Accuracy} = \frac{\text{Number of relevant items in } K}{\text{Total number of items}} \quad (1)$$

Other than that, if we are developing a fully functional system where we fine-tune the vision language model, we can have several evaluation metrics to obtain a detailed understanding of our document QA system’s performance. The metrics are chosen to capture different aspects of answer quality and they are used to run baselines on InstructBLIP and Llava. BLEU [4], METEOR [5], and ROUGE-L [6] provide measures of surface-level overlap between candidate and reference answers, which helps determine whether the correct words and structural patterns are present. CIDEr [7] is used to account for the agreement across multiple references by weighting the relevance of terms with TF-IDF, offering a view of consensus that is especially useful when diverse reference answers exist. In addition, BERTScore [8] and L3Score [1] assess the semantic similarity between the generated and reference answers. BERTScore uses contextual embeddings to match meaning beyond exact word overlap whereas L3Score employs a large language model to estimate answer correctness based on log-likelihood, directly addressing semantic equivalence. Together, these metrics offer a balanced evaluation that covers both lexical similarity and deeper semantic accuracy.

3.4 Loss Function

We fine-tune clip using the Triplet Margin Loss, which works by making the CLIP to bring similar images closer in an embedding space whereas pushing apart mismatching images away from each other in the embedding space. It uses triplet of images namely anchor, positive sample and negative sample. The loss function in equation 2 makes sure the distance between anchor and positive sample is less than the distance between anchor and the negative sample by at least a specified margin denoted by α .

$$\mathcal{L}(a, p, n) = \max(\|f(a) - f(p)\|^2 - \|f(a) - f(n)\|^2 + \alpha, 0) \quad (2)$$

4 Baseline Evaluation and Extensions

4.1 Baseline Experimentation Details

4.1.1 Image RAG Baseline

We chose CLIP as our baseline model for the image retrieval component of our system. We tested the performance of the pre-trained CLIP model on the SPIQA test set. The model had to pick the image based on the question from the all_figures column which contained all the figures in the research paper, where the correct image would be the figure present in the reference_figure column. We observed a 34% accuracy when retrieving 1 image, 60% on 3 images and 75% on 5 images.

4.1.2 Zero-Shot Model Baselines

We chose Gemini 1.5 Pro and InstructBLIP-7B as the baseline models for this project due to their complementary nature. Gemini 1.5 Pro is a closed model and its one of the well known and high performing model of Google that can perform various tasks which makes it a benchmark for evaluating the limits fine-tuned models. On the other hand, InstructBLIP-7B is an open model allowing for fine-tuning. However, we wont be fine-tuning InstructBLIP-7B in this work as the authors of SPIQA have already done it.

By choosing both open and closed model, we ensure a reliable and diverse evaluation and another reason is that both models are known for performing well in multimodal tasks which makes them ideal candidates for establishing baseline in our experiments. For experimentation, we give both caption and image as input for InstructBLIP-7B model whereas for Gemini 1.5 Pro, we give only the

caption as input as we were limited by the token limit and number of requests per day of the free API key.

4.2 Extensions

As discussed in 3.1, in this project, we have fine-tuning CLIP using the SPIQA dataset. Based on the results of SPIQA paper where fine-tuning on open models improved the performance, we can experiment in fine-tuning models such as Qwen-2.5-VL-7B-Instruct if we want to extend the existing work. Other than these, we can fine-tune ColBERTv2 for improved text-based retrieval.

5 Result and Analysis

5.1 Experimentation Results

Our current work focused on Image RAG using a CLIP model, while initial fine-tuning with 10k samples 3 degraded performance on the baseline 2 probably due to catastrophic forgetting, increasing the sample size to 50k 4 allowed the model to learn an improved representation of the underlying distribution leading to improved performance on the test set.

We noticed a sharp reduction in image retrieval performance as the number of images in a research paper increased, the threshold being close to 10 images, as the number of images in a paper crossed 10 the CLIP models ability to retrieve the relevant image deteriorated.

The experiment results in Table 1 show a clear performance difference between InstructBLIP and Gemini 1.5 Pro. Gemini 1.5 Pro obtains higher scores in BLEU metrics (BLEU@1 to BLEU@4), METEOR, ROUGE-L, BERTScore F1, and L3Score. In contrast, InstructBlip achieves a slightly higher score in the CIDEr metric. Overall, the data indicate that Gemini 1.5 Pro performs better on most metrics compared to InstructBlip.

Table 1: Comparison of InstructBlip and Gemini 1.5 Pro Evaluation Results (in %)

Metric	InstructBlip	Gemini
BLEU@1	1.5	22.5
BLEU@2	1.2	15.7
BLEU@3	1.0	12.0
BLEU@4	0.9	9.4
METEOR	5.1	18.4
ROUGE-L	10.3	24.1
CIDEr	33.4	28.8
BERTScore F1	29.3	54.2
L3Score	7.05	24.69

Table 2: Top-k Retrieval Accuracy of Vanilla CLIP Model (in %)

Top-k Metric	Accuracy
Top-1	34.38
Top-3	60.66
Top-5	75.83

Table 3: Top-k Retrieval Accuracy of CLIP Model after fine-tuning with 10,000 samples (in %)

Top-k Metric	Accuracy
Top-1	29.43
Top-3	56.61
Top-5	71.62

Table 4: Top-k Retrieval Accuracy of CLIP Model after fine-tuning with 50,000 samples (in %)*

Top-k Metric	Accuracy
Top-1	~38
Top-3	~75
Top-5	~85

*The results mentioned in Table 4 are approximate values. The model was fine-tuned on PSC, and due to a tornado, PSC went down. So we were unable to access the final and accurate results.

5.2 Discussion

We believe that using a larger VL model for Image-RAG would lead to significant improvements in top 1 retrieval as these models are trained on a larger data volume and have a larger capacity to better differentiate nuances.

While OCR could be a useful method for extracting knowledge from images of tables, it would not be useful in extracting meaningful information from schematic and scientific diagrams.

We believe that future-work on sparse retrieval to enhance long-context abilities of smaller open-source LLMs could be a meaningful pursuit in the future.

6 Future Directions

A remaining challenge in long-document question answering is that a model may lose track of relevant information when the input grows beyond its native context window. To address problem, one future exploration direction is combining dual chunk attention and sparse attention. Dual chunk attention addresses the context length issue by breaking the input into equal-sized segments and then running coordinated attention passes [11]. Because each pass works on limited slices or summaries rather than the full sequence, computation and memory grow roughly in proportion to sequence length rather than its square, letting the model handle much longer texts without extra training.

An alternative approach is sparse attention, and in particular the MInference method for the prefill stage of generation [12]. Rather than letting each query attend to every key, by skipping most of the dense attention matrix, MInference cuts latency by up to an order of magnitude in the prefill phase while keeping the model’s output quality on benchmarks such as InfiniteBench and PG-19 nearly unchanged. This makes it possible to extend existing language models to much longer contexts without changing their weights.

We have attempted to include this approach in this project; however, due to two reasons, we ultimately chose to not include it. The primary reason is that, we realized that this long context problem is out of the scope of discussion of this project, as the current SPIQA dataset does not show any evidence of being suffered from context length insufficiency. The secondary reason is that, due to issues of the AWS EC2 Ubuntu kernel, CUDA 12.1 cannot be successfully installed. Therefore, even though we have found and incorporated a supposedly working implementation, we are unable to test and tune the parameters, within the time constraint of this project.

7 Conclusion

In conclusion, we address the research gap in image retrieval for multimodal QA systems for research papers, where LLMs must be provided the relevant images to reason upon based on user query. For this we fine-tuned a CLIP model that showcased a 4% improvement on top-1 retrieval, 15% improvement on top-3 retrieval and a 10% improvement on top-5 retrieval. We demonstrate that improvements in image retrieval will enable the use of smaller open-source LLMs that have shorter context lengths, and could also be valuable for larger LLMs with longer context lengths as they suffer from lost in the middle where they hallucinate on context provided in the middle of the user prompt reducing performance of In-context Learning techniques.

Acknowledgments

We would like to express our sincere gratitude to Carnegie Mellon University for providing the resources that will make this project possible. We are thankful to Professor Bhiksha Raj for his guidance throughout the course 11-785: Introduction to Deep Learning, which provided the foundation for this work. Our appreciation also goes to our mentor, Michael Kireeff whose insights and support were instrumental in the completion of this project.

References

- [1] Pramanick, S., Chellappa, R., & Venugopalan, S. (2024). SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. ArXiv (Cornell University).
- [2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, & Haofen Wang. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey.
- [3] Brandon McKinzie, Vaishal Shankar, Joseph Cheng, Yinfei Yang, Jonathon Shlens, & Alexander Toshev. (2023). Don't Fine-Tune, Just Retrieve: Making Multimodal Transformers More Robust.
- [4] Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics.
- [5] Lavie, A., & Agarwal, A. (2007). Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 228–231). Association for Computational Linguistics.
- [6] Lin, C.Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out (pp. 74–81). Association for Computational Linguistics.
- [7] Ramakrishna Vedantam, C. Lawrence Zitnick, & Devi Parikh. (2015). CIDEr: Consensus-based Image Description Evaluation.
- [8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, & Yoav Artzi. (2020). BERTScore: Evaluating Text Generation with BERT.
- [9] google, GitHub - google/spiqa: Code release for 'SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers' [NeurIPS D&B, 2024], GitHub, 2024. <https://github.com/google/spiqa>.
- [10] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, & Zihan Qiu. (2025). Qwen2.5 Technical Report.
- [11] Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, Lingpeng Kong. (2024). Training-Free Long-Context Scaling of Large Language Models.
- [12] Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, Lili Qiu. (2024). MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention.
- [13] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv.org, Feb. 26, 2021. <http://arxiv.org/abs/2103.00020>

[14] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, “ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction,” arXiv.org, Jul. 10, 2022. <https://arxiv.org/abs/2112.01488>

Team Contribution

Anirudh Joshi

1. Ran Baseline on InstructBLIP and evaluated on BLEU, METEOR, ROGUE, CIDEr and BERTScore F1
2. Ran baseline using CLIP
3. Worked on Image-based RAG (CLIP)

Manigandan Ramadasan

1. Ran Baseline on Gemini and evaluated on BLEU, METEOR, ROGUE, CIDEr and BERTScore F1
2. Worked on Text-based RAG (ColBERTv2)
3. Developed script for zero-shot classification using Qwen

Yiwei Chen

1. Evaluated InstructBLIP and Gemini on L3Score
2. Tried implementing Sparse Attention