ColFlor: Towards BERT-Size Vision-Language Document Retrieval Models

Ahmed Masry
York University
Ontario, Canada
masry20@yorku.ca

Enamul Hoque
York University
Ontario, Canada
enamulh@yorku.ca

Abstract

Traditional document retrieval systems for PDFs, charts, and infographics rely heavily on Optical Character Recognition (OCR) pipelines to extract textual content, a process that is both error-prone and resource-intensive. Recent advancements in multimodal models like ColPali have enabled OCR-free retrieval by processing documents directly as images, but their large size (three billion parameters) makes them computationally expensive and impractical for large-scale applications. To address this limitation, we introduce ColFlor, an efficient OCR-free visual document retrieval model with only 174 million parameters. ColFlor achieves comparable performance to ColPali on text-rich English documents—with only a 1.8% decrease in performance (measured by NDCG@5 metric)—while being significantly faster in image encoding (5.25 times faster) and query encoding (9.8 times faster). This makes OCR-free document retrieval systems more cost-effective for large-scale applications and more accessible to users with limited computational resources.

1 Introduction

Information retrieval (IR) systems play a vital role in real-world applications, powering search engines to find relevant information on the web and enabling efficient document retrieval from large databases. Over recent years, significant advancements in IR systems have leveraged machine learning models [8, 5, 10] to rank and retrieve information based on their relevance to the user's queries. However, these models are language-based and can not directly process information embedded in visual format such as PDFs, charts, or infographics. To overcome this limitation, traditional approaches rely on Optical Character Recognition (OCR) techniques to extract textual content from visual documents for use in information retrieval systems. However, OCR-based methods are often computationally intensive, costly, and error-prone, particularly for documents with complex layouts. Recent advancements in vision-language models (VLMs) [1, 11, 7, 2] have revolutionized document retrieval by enabling OCR-free systems. These models directly process document images, eliminating the need for OCR and mitigating the associated errors and computational overhead. Leading OCR-free models like ColPali [4] have demonstrated state-of-the-art performance on multimodal document retrieval tasks by leveraging VLMs such as PaliGemma [1] as their backbone. Despite their effectiveness, the massive size of these models limits their practicality, especially in resource-constrained environments or large-scale retrieval applications where latency is critical. To address these challenges, this paper introduces ColFlor, an efficient OCR-free document retrieval model. At 174 million parameters, ColFlor is 17 times smaller than ColPali, achieving competitive performance while delivering substantial speed gains. It is 9.8 times faster in query encoding and 5.25 times faster in image encoding, with only a 1.8% reduction in performance on text-rich English documents. This efficiency makes ColFlor a viable alternative to larger models and a practical solution for large-scale applications and resource-limited environments. Our codebase, model weights, and an interactive demo are available at: `https://github.com/AhmedMasryKU/colflor`.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

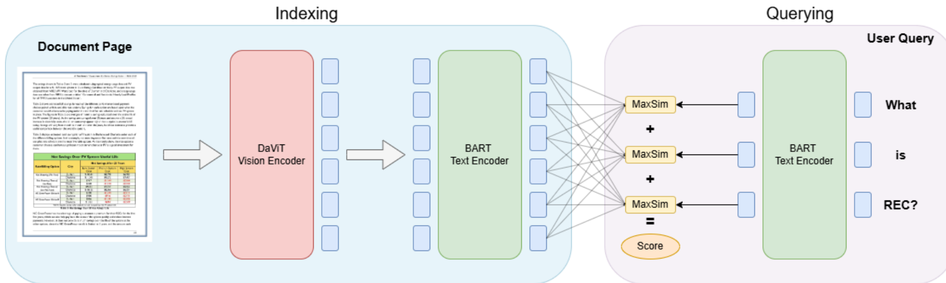

Figure 1: The ColFlor model architecture, showcasing the vision encoder, text encoder, and late interaction retrieval mechanism.

2 Model Architecture

ColFlor builds on the Florence-2 architecture [12], leveraging its vision and text encoders while discarding the text auto-regressive decoder. The model architecture follows a two-stage process of indexing and querying, as shown in Figure 1.

Indexing: During the indexing phase, the DaViT vision encoder [3] extracts visual features from input document images, transforming them into a sequence of visual embedding vectors. These embeddings are then processed by a BART-based text encoder [6], generating rich contexualized embeddings for the document. To reduce storage requirements, these contextualized embeddings are projected into 128-dimensional vectors using a linear layer, similar to the techniques employed in ColBERT [5] and ColPali [4].

Querying: In the querying phase, the text encoder processes the user query to produce query embeddings. These embeddings are then compared to the stored document embeddings using the MaxSim operation, a late-interaction retrieval mechanism [5]. Unlike traditional retrieval methods, which reduce documents and queries into single vector representations [10, 9], MaxSim computes fine-grained similarities between the bags of contextualized embeddings, preserving the detailed semantic structure of both the query and the document.

3 Training Setup

We initialized the model weights from the Florence-2-base model, with the exception of the new linear projection layer, which was randomly initialized. Initially, training was unstable, and the loss failed to converge despite doing some hyperparameter search. To address this, we first removed the randomly initialized projection layer and trained the model for 5 epochs. This stabilized the training and improved convergence. Afterward, we reintroduced the linear projection layer and fine-tuned the model for 40 epochs on the ViDoRe dataset [4], using a learning rate of 2e-5 and a batch size of 64 on 4-A100 GPUs.

4 Evaluation

Performance: We evaluated ColFlor using the NDCG@5 metric on the ViDoRe benchmark [4], which consists of 10 subcategories of document retrieval tasks. We group them as follows:

- **Text-rich English Documents**: Includes academic datasets like DocVQA, TatDQA, and real-world practical data like AI, Energy, Government Reports, and Healthcare.
- **Figure Documents**: Includes InfoVQA and ArxivQA, which primarily consist of complex visuals such as figures, diagrams, and infographics.
- **French Documents**: Includes TabFQuAD and Shift, testing the model's multilingual capabilities.

As shown in Table 4, ColFlor performs comparably to ColPali on text-rich English documents, with only a 1.8% decrease in the average performance, despite its significantly smaller size. Notably, ColFlor outperforms ColPali on TatDQA, a VQA dataset derived from publicly available real-world

	Text-rich English Documents							Figures			French Documents		
	DocQ	TATQ	AI	Energy	Gov.	Health.	**Avg.**	InfoQ	ArxivQ	**Avg.**	TabF	Shift	**Avg.**
SigLIP (Vanilla)	30.3	26.2	62.5	65.7	66.1	79.1	55.0	64.1	43.2	53.7	58.1	18.7	38.4
BiSigLIP (+fine-tuning)	32.9	30.5	74.3	73.7	74.2	82.3	61.3	70.5	58.5	64.5	62.7	26.5	44.6
BiPali (+LLM)	30.0	33.4	71.2	61.9	73.8	73.6	57.3	67.4	56.5	61.9	76.9	43.7	60.3
ColPali (+Late Inter.)	**54.4**	65.8	**96.2**	**91.0**	**92.7**	94.4	**82.4**	**81.8**	**79.1**	**80.5**	**83.9**	**73.2**	**78.6**
ColFlor (Ours)	51.06	**66.2**	90.97	88.43	91.2	**95.95**	80.63	65.49	69.86	67.67	43.48	25.37	34.42

Table 1: Performance comparison of ColFlor against state-of-the-art OCR-free retrieval models on the ViDoRe benchmark across different categories. Metrics are reported as NDCG@5, with ColFlor demonstrating competitive performance despite its significantly smaller model size.

financial reports, as well as the Health dataset. This highlights ColFlor's potential for real-world applications and its ability to scale efficiently. The performance gap is more pronounced in the Figures category, likely due to the backbone model's (Florence-2) focus on text-rich documents and limited training on figures. We plan to address this by continuing the pretraining of Florence-2 on figures before finetuning it on the document retrieval task in the future. Lastly, ColFlor performs poorly on French documents, as Florence-2 was designed for English only and lacks multilingual support.

Efficiency: The ColFlor model aims to offer an efficient, affordable, yet high-performing alternative to ColPali, making the new OCR-free document retrieval paradigm accessible to users with limited computing resources. We benchmarked both models' forward passes on a T4 GPU using the float32 data type. For image encoding, we used a batch size of 32 for ColFlor and 2 for ColPali. For query encoding, we used a batch size of 1 to simulate online querying. Our experiments show that ColFlor is 5.25 times faster for image encoding and 9.8 times faster for query encoding. Additionally, ColFlor processes images at a higher resolution (768x768 vs. 448x448 for ColPali) while producing fewer contextualized embeddings (587 vs. 1024) which reduces the embeddings storage costs.

5 Conclusion

We introduced ColFlor, a BERT-sized model for OCR-free document retrieval. ColFlor is significantly smaller than ColPali and provides much faster image and query encoding, while maintaining nearly the same performance on text-rich English documents. Future work includes further training of the backbone model, Florence-2, on figure datasets to enhance figure understanding, as well as developing a multilingual variant to broaden ColFlor's application scope and support diverse languages.

Acknowledgements

This research was supported by the Natural Sciences Engineering Research Council (NSERC) of Canada and Canada Foundation for Innovation (CFI). Additionally, it received support through a GCP credits award from Google's PaliGemma Academic Program.

References

[1] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024.

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024.

[3] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers, 2022.

[4] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2024.

[5] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert, 2020.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[8] Bhaskar Mitra and Nick Craswell. An updated duet model for passage re-ranking, 2019.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[12] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023.