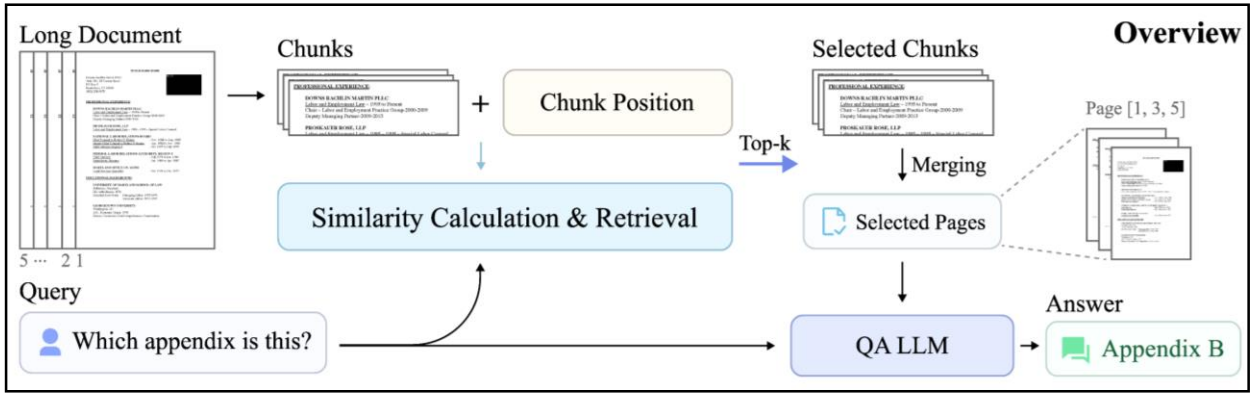
Position-aware Modeling for Fine-grained Visually-rich Long Document Understanding

Yixiao MA, Shulan RUAN, Zijie SONG, Xin ZHANG, Yuze ZHAO, Zhenya HUANG, Enhong CHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-026-51131-x

Problems & Ideas

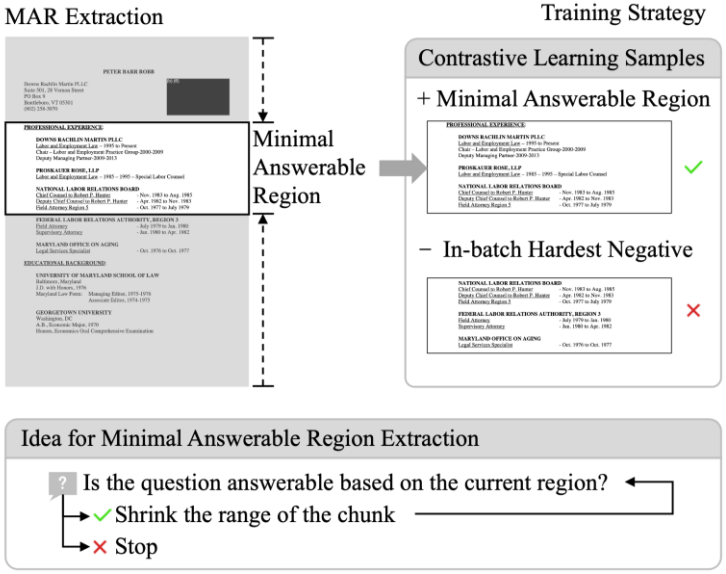
- Problems of visually-rich long document understanding:
  - Requires identifying query-relevant information across multiple pages while preserving layout and positional cues.
  - Existing methods either process entire documents inefficiently or lack fine-grained position-aware retrieval, leading to noise from irrelevant content and reduced accuracy.
- Ideas: A position-aware retrieval framework that leverages fine-grained document chunks together with their corresponding positional information to retrieve relevant content.



Overview of the position-aware fine-grained multi-page document retrieval–QA framework. Each document page is segmented into fine-grained chunks, whose visual content is jointly encoded with their corresponding positional information to compute relevance to the query. The QA model then answers the question based on the pages associated with the top-k most relevant document chunks.

Main Contributions

- Contributions:
  - A novel position-aware fine-grained retrieval–QA framework that jointly leverages document chunks and their positional cues to enable more accurate content retrieval in visually-rich long documents;
  - A minimal answerable region based training strategy for the retriever, substantially improving training efficiency and significantly reducing the amount of required training data;
  - Fine-grained retrieval reduces the number of pages the QA model needs to process, significantly saving resources.



MAR extraction and its use in training, serving as positive samples in contrastive learning for stronger convergence.

Model	#Trainable Param.	#Train Data	Epoch	ANLS
ColPali†	39.29M	~119K	1	0.78
SV-RAG†	25.37M	>1000K	4	0.78
PDU	**9.64M**	**~35K**	**1**	**0.80**

PDU achieves competitive QA accuracy with significantly less training data.

Model	Avg. Pages↓	QA Time↓	QA Mem.↓	ANLS↑
ColPali†	5.00	1.5	41.23	0.78
SV-RAG†	5.00	1.5	40.94	0.78
PDU	**3.74**	**1.1**	**32.35**	**0.80**

PDU enables the QA model to process fewer pages while maintaining competitive accuracy.