テキスト抽出不要のRAGを実現するColPali

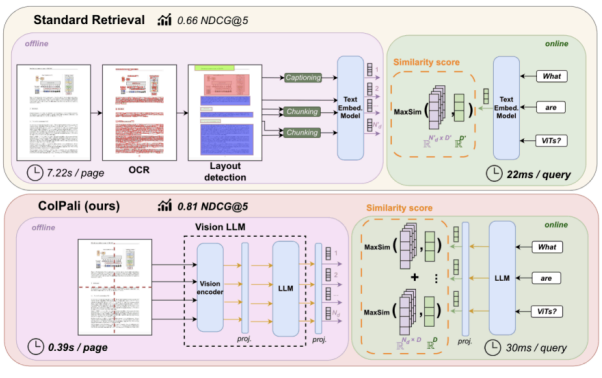
2024/07/30に公開

PDF image LLM embedding RAG Tech

導入

こんにちは、株式会社ナレッジセンスの須藤英寿です。普段はエンジニアとして、LLMを使用したチャットのサービスを提供しており、とりわけRAGシステムの改善は日々の課題になっています。

本記事では、画像の情報をそのままベクトルデータにして検索する手法、ColPaliについて解説します。



sasakuna

フォロー

Knowledge Sense, Inc. CTO

目次

Zennからのお知らせ

サマリー

通常、RAGでは文書データからテキストを抽出して、その文字をベクトルデータに変換します。しかしColPaliは、文書データを画像として認識してベクトル化を行います。画像として保管することでテキスト化できない情報を扱うことができます。他にもベクトルを複数に分解することで精度を改善し、テキストの抽出が必要ないことからデータ保管時のコストの大幅な低減などのメリットを享受できます。

PDFのデータを保管する際には、ColPaliモデルに正規化したPDF画像を入力として渡し1024個の128次元ベクトルを生成します。検索時には、クエリごとにColPaliモデルのテキストエンコーダーに入力し、類似度の合計値の高いPDFのページデータを取得します。

▼ 問題意識

PDFのRAG利用の難しさ

一般的なRAGでは、文書内のテキストデータを細かい単位で切り分けてベクトルデータベースに保管されます。PDFの場合ファイルからテキストデータを抜き出す必要があるのですが、実はこの工程がかなり難しいタスクで、精度を出すためにもかなりの計算時間を要します。これはPDFファイルのデータの保

持の方法に起因するため、改善の難しいものとなっています。

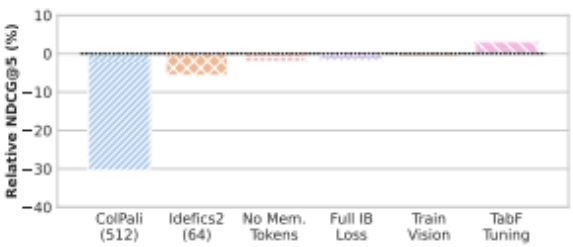
手法

ColPaliは、画像データを1024個のパッチに分割しそれぞれをメタデータと結びつけた状態でベクトルデータベースに保管します。検索の際にはテキストをトークン化し、それぞれのColPaliのテキストエンコーダに入力。そしてベクトルデータの類似度を元に検索し、合計点が最も高い画像を順に取得します。

これらの手順の詳細は以下のようになります。

事前準備

ColPaliでは独自のEmbedingモデルを使用します。PaliGemma-3Bモデルをベースに、1024個の128次元ベクトルデータを出力するように調整しています。このモデルは公開されており、ここからさらなるファインチューニングを行っても精度として大きな変化はないと言及されています。



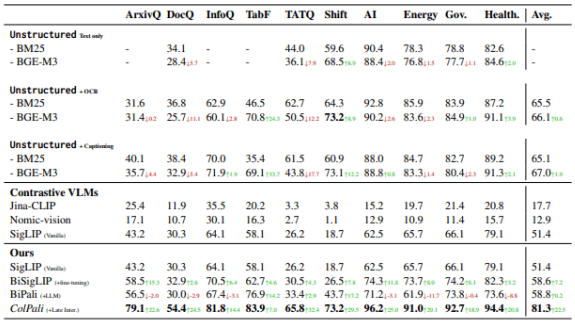
保存方法

1. PDFの各ページごとに画像に変換
2. PDFを入力サイズに合わせてリサイズ
3. ColPaliモデルを使用して、ページごとに1024個の128次元ベクトルに変換
4. すべてのベクトルデータをPDF本体とページ番号のメタデータと結びつけて保存

検索方法

1. 検索クエリ用の文字列をトークンに変換
2. ColPaliモデルのテキストエンコード機能を利用して、ベクトルデータをトークンごとに作成
3. ColBertを使用してトークンごとに類似度の高いベクトルを検索、ページ単位でスコアを集計
4. スコアの高い順にPDFのページを並び替える

成果

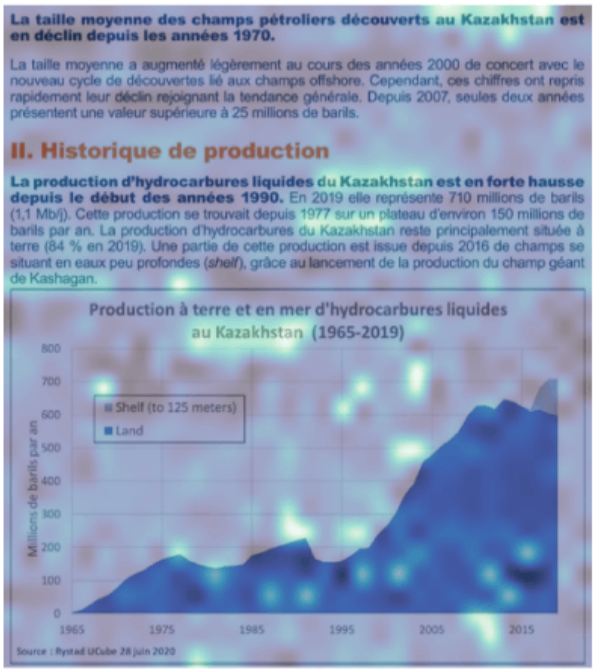


	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured Text only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4	-	-	36.1	68.5	88.4	76.8	77.7	84.6	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4	25.7	60.1	70.8	50.5	**73.2**	90.2	83.6	84.9	91.1	66.1
Unstructured + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	35.7	32.9	71.9	69.1	43.8	73.1	88.8	83.3	80.4	91.3	67.0
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP (+fine-tuning)	58.5	32.9	70.5	62.7	30.5	26.5	74.3	73.7	74.2	82.3	58.6
BiPali (+LLM)	56.5	30.0	67.4	76.9	33.4	43.7	71.2	61.9	73.8	73.6	58.8
ColPali (+Late Inter.)	**79.1**	**54.4**	**81.8**	**83.9**	**65.8**	**73.2**	**96.2**	**91.0**	**92.7**	**94.4**	**81.3**

- ArxivQ DocQ InfoQ TabF TATQなどの既存のベンチマーク手法すべてで、従来手法以上の成果
- Shift AI Energy Gov. Healthなどの独自に作成した、より現実のタスクに即したベンチマーク手法であっても従来手法以上の成果

Model	Embedding size (KB)
BGE-M3	8.60
BM25 (dense emb.)	3.00
BM25 (sparse emb.)	**1.56 ± 0.51**
ColPali (float16)	256

- Embeddingの保存サイズは従来の手法と比較して圧倒的に容量が大きい
- 最適化の余地があり最終的には数倍程度の差に収まるとしている



- パッチ情報との類似度を元に画像内のどの場所に着目したかを可視化できる

まとめ

ColPaliは、PDFの検索に焦点を当てていましたが、実際には画像全般に応用が利くので非常に使いやすい手法かなと思います。加えてベクトル化するスピードも非常に早く、これはPDFをRAGで利用するための既存の手法と比較すると大きなアドバンテージかと思います。

また個人的に特に面白いと感じたのは、一つのページを1024個のベクトルデータ

に分解するという点です。通常テキストデータをベクトル化する際には、一つのチャンクに対して一つのベクトルデータを割り当てます。Embeddingモデルもその想定で学習を進めているので基本的にそれで問題なく精度が出せると思います。一方でクエリのコンテキストごとに、テキストデータの持つ意味というのはすこしずつ変わって来るはずなので、本当に最適化するのであればそういった想定も必要なのではと考えていました。そういった点で、一つの画像を1024個のベクトルデータに変換しそれぞれが別々の意味を持つというのは応用が利きそうな考え方に感じます。

話をColPaliに戻して、この手法で気になった点としてはEmbeddingを用いた検索の一般的な弱点である、固有の単語に弱いという問題が解消しづらいことです。ただしこの点は、PDFの解析時に構造を把握せず単語だけ把握すれば良くなった、と捉えてBM25等の手法とハイブリッド検索することで弱点を補えるのではと考えています。

PDFをRAGで使ってもなかなか成果が出ない。PDFのデータ保存に時間がかかって困っている。という方は一度利用を検討してみても良いのかなと思います。

sasakuna

Knowledge Sense, Inc. CTO

フォロー

ナレッジセンス - 生成AIとRAGの実装戦略・技術ブログ

Publication

株式会社ナレッジセンスは、「大企業の知的活動を最速にする」をミッションに掲げ、社内データ検索ができるAIチャットボットを開発・提供しているスタートアップです。このブログでは、LLMや検索技術、RAGの実装戦略について知見を共有します。生成AIやRAG技術を使って最高品質の実装をしたいエンジニア向けのコンテンツです

フォロー

Discussion

ログインするとコメントできます

Login

Read next

APIをそのままMCPサーバーにするな


いばら… in NCDC テックブロ…
4日前 69

「手作り RAG システム」で RAG の仕組みを学び直す


Etsuji Na… in Google Cloud Ja…
3日前 54

AMD Ryzen AI Max+ 395で gpt-oss-20bを動かすためのAPIサーバーとモデルの評価


anjn 7時間前

VRAMの限界を突破する？次世代技術「vLLM」の衝撃


あやと@AI自動… 23時間前 2

DifyのRAG検索を廃止して独自検索APIを構築した話


inori 27分前

「1億語コンテキスト」は“理論上もう届く” — 現場で何を変えるべきか


宝宝 (baobao) 9時間前 1