VILE: Block-Aware Visual Enhanced Document Retrieval

Huaying Yuan
Zhicheng Dou
hyyuan@ruc.edu.cn
dou@ruc.edu.cn
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China

Yujia Zhou
Yu Guo
zhouyujia@ruc.edu.cn
yu_guo@ruc.edu.cn
School of Information
Renmin University of China
Beijing, China

Ji-Rong Wen
jrwen@ruc.edu.cn
Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, China
Renmin University of China
Beijing, China

ABSTRACT

Document retrieval has always been a crucial problem in Web search. Recent works leverage pre-trained language models to represent documents in dense vectors. However, these works focus on the textual content but ignore the appearance of web pages (e.g., the visual style, the layout, and the images), which are actually essential for information delivery. To alleviate this problem, we propose a new dense retrieval model, namely VILE, to incorporate visual features into document representations. However, because a web page is usually very large and contains diverse information, simply concatenating its textual and visual features may result in a cluttered multi-modal representation that lacks focus on the important parts of the page. We observe that web pages often have a structured content organization, comprising multiple blocks that convey different information. Motivated by the observation, we propose building a multi-modal document representation by aggregating the fine-grained multi-modal block representations, to enable a more comprehensive understanding of the page. Specifically, we first segment a web page into multiple blocks, then create multi-modal features for each block. The representations of all blocks are then integrated into the final multi-modal page representation. VILE can better model the importance of different content regions, leading to a high-quality multi-modal representation. We collect screenshots and the corresponding layout information of some web pages in the MS MARCO Document Ranking dataset, resulting in a new multi-modal document retrieval dataset. Experimental results conducted on this dataset demonstrate that our model exhibits significant improvements over existing document retrieval models. Our code is available at https://github.com/yhy-2000/VILE.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

KEYWORDS

Document retrieval; Multi-modal; Visual

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom.

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615107

ACM Reference Format:
Huaying Yuan, Zhicheng Dou, Yujia Zhou, Yu Guo, and Ji-Rong Wen. 2023. VILE: Block-Aware Visual Enhanced Document Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3615107

1 INTRODUCTION

Document retrieval [23, 33], aiming to extract relevant documents from a large collection based on user queries, is a critical task in information retrieval. It serves as a fundamental component in various downstream tasks, including open-domain question answering, document ranking, retrieval-augmented generation, etc. Among all the methods, dense retrieval is a prominent document retrieval approach [13, 35, 38]. Empowered by pre-trained language models (PLMs) [5, 19, 36], dense retrievers utilize a siamese architecture to encode queries and documents into a semantic space, facilitating efficient similarity computation between them.

While existing dense retrieval models have made remarkable strides in capturing document semantics, they exhibit a limitation in effectively representing web pages. These models predominantly rely on textual content, thereby neglecting the rich and comprehensive visual information presented on web pages. Visual information presented in web pages, such as visual elements (e.g., images and graphics), display styles (e.g., colors and font sizes), and content layouts (e.g., headers and sidebars), provides valuable supplementary insights that complement the textual content [11, 14, 32]. By disregarding these visual cues, existing models fail to capture the holistic nature of web pages and miss crucial information. Thus, there is a clear need to integrate visual information into the document representation process to overcome this limitation and achieve a more comprehensive understanding of web pages.

Prior research has explored the integration of visual information into web search, as evidenced by works such as VIP [7] and VITOR [34]. These studies have highlighted the benefits of incorporating visual information into document features, resulting in improved web search outcomes. However, these approaches primarily focus on extracting visual features in isolation, which are then simply concatenated with textual features at the final layer of the model. This coarse-grained integration fails to capture the intricate interactions between visual and textual features. Consequently, in this paper, we propose a method that aims to incorporate visual features into document representations with the objective of further exploring the potential of visual information in web search. Our

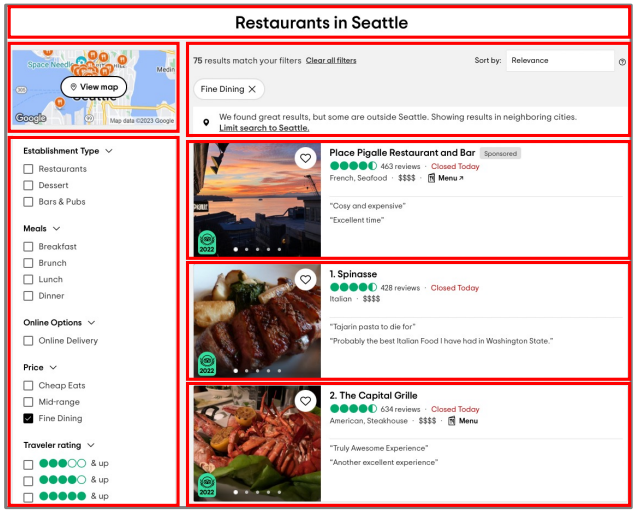

Figure 1: Illustration of content blocks contained in a web page. There are many distinct blocks in the web page, comprising a header, a search block, a map, a sidebar, and three content blocks with bolded titles and illustrated figures.

proposed methodology seeks to overcome the limitations of previous approaches by enabling a more fine-grained and comprehensive fusion of visual and textual features.

Factually, web pages usually encompass complex structures with diverse elements. Figure 1 illustrates a snapshot of a typical web page. The page consists of various content blocks, including a header, a search block, a geographical map, a sidebar, and three record blocks containing detailed restaurant information. The record blocks are particularly noteworthy as they feature bolded titles and illustrated figures, which enhance their visual prominence and distinguish them from the surrounding elements. These blocks serve as focal points within the web page, containing valuable information that is pertinent to the overall page content and user engagement. We argue that understanding and effectively representing these visual content blocks are crucial to the retrieval task, and we should construct a multi-modal representation model that can effectively aggregate the finer-grained block representations.

In this paper, we present **VILE** (Block-aware **VI**sua**L** **E**nhanced Dense Retriever), a new dense retriever designed to enhance the learning of comprehensive web page representations. The proposed method consists of three main steps. **Firstly**, the textual content and screenshots of the entire web page are transformed into text embeddings and image representations, respectively. These representations are then fed into a multi-modal page transformer, enabling effective interaction between visual and textual features and yielding a more holistic document representation. **Secondly**, we focus on capturing finer-grained multi-modal semantics and emphasize learning the significance of crucial blocks. A page segmentation module is employed to partition the web page into blocks. Each block undergoes the same process as the entire web page, resulting in multi-modal representations that encompass more comprehensive information. **Thirdly**, by integrating these multi-modal block representations, VILE generates a more comprehensive representation of the entire web page.

Due to the requirement of our method on multi-modal information, we construct a novel multi-modal document retrieval dataset based on the MS MARCO Document Ranking dataset [24], which collects screenshots and corresponding layout information of web pages. Experimental results on this dataset proved the effectiveness of VILE in the document retrieval task.

In conclusion, the main contributions of this paper are summarized as follows: (1) We introduce a new approach that enhances document representation by leveraging visual information. This integration of visual features enables a more comprehensive and enriched representation of the documents. (2) We propose a method to effectively capture the varying importance of different content regions within the documents, resulting in a high-quality multi-modal representation that emphasizes crucial sections. This approach improves the overall quality of the document representation. (3) We create a multi-modal document retrieval dataset. Experimental results conducted on this dataset demonstrate the superiority of our proposed model over existing document retrieval models, underscoring its effectiveness and potential for practical applications.

2 RELATED WORK

2.1 Dense Retrieval

Dense Retrieval (DR) is an approach that leverages embeddings to represent queries and documents. In the offline stage, DR encodes documents with dense representations and constructs the document index. During the online stage, it encodes input queries and conducts similarity search [13, 17, 35, 38]. Dense representations, as opposed to sparse bag-of-words representations like BM25 [29], retain a greater amount of semantic information and demonstrate superior performance in online search tasks. Current studies mainly focus on developing enhanced methods for negative sampling to train the document encoder. For instance, DPR [13] introduced the concept of in-batch negatives for training, while ANCE [35] employs a learning mechanism that globally selects static hard negatives from the entire corpus. ADORE [38] presents a hybrid approach combining random negatives and static hard negatives. With the advancements in pre-training language technology [2, 5], there has been a growing interest in incorporating search-oriented tasks [1, 8, 9] during the pre-training process, such as RetroMAE [20]. Although these approaches have shown effectiveness, they might not sufficiently capture the representation of web pages in real-world search engines. In practice, web pages possess distinctive structures and visual patterns that contribute to their semantics and influence users' browsing experiences. Therefore, it is imperative to incorporate the visual information of web pages into the document representations to ensure more comprehensive semantic modeling.

2.2 Visual Search

Several studies have made attempts to integrate visual features into information retrieval. For instance, VIP [7] argues that visual information derived from the layout of web pages holds value for relevance modeling in web search. They propose the utilization of web screenshots as a complementary input for Learning To

Rank (LTR) features [27]. VITOR [34] introduces synthetic saliency heatmaps that explicitly capture how users view web pages. These approaches primarily rely on coarse-grained features such as BM25, TF-IDF, and pagerank score as content features, without considering the interactions between visual and textual features. Zhang et al. [40] attempts to incorporate structural information and visual information of search engine result pages into document ranking, which achieves better performance than other ranking solutions and the original ranking result of search engines. This work highlights the importance of both visual and structural features in web search. In comparison to these existing studies, our research focuses on more fine-grained block-aware interactions between the textual and visual modalities.

3 METHODOLOGY

As mentioned in Section 1, existing text-only document representation models are insufficient to capture the abundant visual information embedded within web pages. In this paper, we propose a visually enhanced dense retriever, namely VILE, which effectively integrates visual information for document retrieval. The overview of VILE is shown in Figure 2. Firstly, the textual content and screenshots of the full web page are transformed into text embeddings and image representations, respectively. These representations serve as inputs to a multi-modal page transformer, facilitating the effective interaction between visual and textual features and resulting in a more holistic document representation. Furthermore, to strengthen the semantics of vital blocks and capture finer-grained multi-modal semantics, we consider individual web page blocks. A page segmentation module is utilized to partition the web page into blocks. Each block follows the same process as the entire web page, yielding multi-modal representations but with more comprehensive information. By concatenating these multi-modal block representations, the multi-modal page transformer derives a more comprehensive representation of the entire web page. In the subsequent sections, we will provide detailed explanations of these components.

3.1 Task Definition

Given a query q, the target of document retrieval is to recall the top k most relevant documents from a large collection C. Recent advanced works mainly employ a dual architecture to encode the query q and a document d to embeddings and use the similarity function to compute the ranking scores, denoted as:

$$\mathcal{R}(q, d) = \langle \vec{X}_{q;\phi}, \vec{X}_{d;\phi} \rangle, \tag{1}$$

where $\vec{X}$ represents embeddings encoded by the dense retriever with parameters ϕ, such as BERT and RoBERTa [5, 19]. And $\langle , \rangle$ is the similarity function. For existing text-only models, the document representation is learned from the text of the document, the document representation can be denoted as:

$$\vec{X}_{d;\phi} = f_{\phi}(d_{\mathrm{T}}), \tag{2}$$

where $f(\cdot)$ is a text-base encoder, and d_{T} is the text information of document d. These works overlook the substantial significance of visual information in semantic understanding. To learn more abundant representations for real-world web pages, we attempt to integrate visual information into the document representation, denoted as:

$$\vec{X}'_{d;\phi} = f'_{\phi}(d_{\mathrm{T}}, d_{\mathrm{V}}), \tag{3}$$

where $f'(\cdot)$ is a multi-modal encoder, and d_{V} is the visual information of the web page. Nonetheless, due to the typically extensive nature of web pages, simply concatenating their textual and visual attributes can lead to a cluttered multi-modal representation lacking emphasis on the vital elements of the page. It is evident that web pages frequently exhibit organized content structures, consisting of various blocks that convey distinct information. This observation serves as the motivation for our proposition: constructing a multi-modal document representation through the consolidation of finely detailed multi-modal block representations. This approach aims to facilitate a more comprehensive comprehension of the page, encompassing its diverse components and facilitating a nuanced understanding of its content. The enhanced representation of a multi-modal document with block information can be expressed as:

$$\vec{X}''_{d;\phi} = f'_{\phi}(d_{\mathrm{T}}, d_{\mathrm{V}}, d_{\mathrm{B}}), \tag{4}$$

where d_{B} refers to the fine-grained block-level multi-modal representations. Incorporating these multi-modal block representations enables the model to capture the semantic content of the web page more effectively, thereby improving retrieval performance.

3.2 Text Representation

For text modality, we follow the text-based dense retrieval methods and utilize the embedding layer of a pre-trained language model, such as BERT [5], to encode words into high-dimensional embeddings. After the embedding matrix, the obtained text information of the web page can be denoted as $d_{\mathrm{T}} = \{d_{\mathrm{T}_1}, d_{\mathrm{T}_2}, \cdots, d_{\mathrm{T}_n}\}$, where d_{T_i} denotes the i-th token contained in the page.

3.3 Incorporating Visual Representation

3.3.1 Visual Representation. Previous works have demonstrated how visual information can enhance information retrieval [7, 34]. However, these models independently extract visual features and just concatenate them with text features at the final layer of the model. Such a coarse-grained manner neglects interactions between visual and textual features, which limits the potential of visual information in document representation.

With the objective of investigating the potential of visual features in the semantic understanding of web pages, we leverage the semantic understanding capabilities of the Vision Transformer [6] to extract visual features from screenshots of web pages. The Vision Transformer employs a transformer architecture to divide images into fixed-size patches and extract their semantic content. The layers of the visual neural network learn varying levels of semantic information. Shallow-level representations provide foundational visual features and assist in component localization, while deep-level representations enhance the comprehension of web page layouts and semantics. By combining both shallow-level representations and deep-level representations, the Vision Transformer becomes proficient in comprehending and processing the multi-grained visual information within web page screenshots. The multi-grained visual representations d_{V} can be represented as

$$d_{\mathrm{V}} = [d_{\mathrm{V}_1}, \cdots, d_{\mathrm{V}_k}], \quad d_{\mathrm{V}_i} = \mathrm{FFN}\left(\text{Visual-Enc}_i(S_d)_{[\mathrm{CLS}]}\right), \tag{5}$$

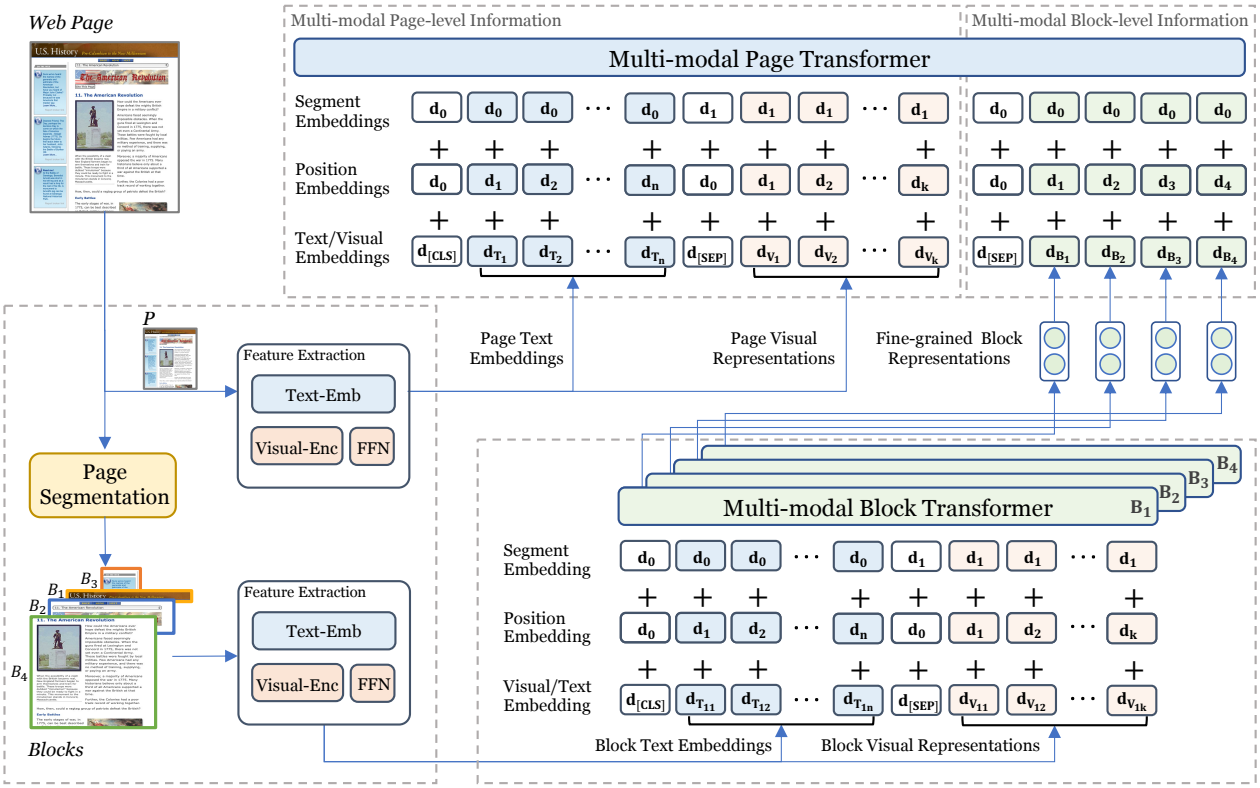

Figure 2: The architecture of VILE. Firstly, the textual content and screenshots of the entire web page are transformed into text embeddings and image representations respectively, which serve as inputs to a multi-modal page transformer. To capture fine-grained semantics, a page segmentation method is employed to partition the web page into individual blocks. Each block undergoes the same process as the entire web page, generating finer-grained multi-modal representations that are specific to that block. By incorporating both textual and visual information, as well as considering the fine-grained block-level representations, multi-modal page transformer facilitates a more comprehensive document representation.

where S_d represents the screenshot of the document d, Visual-Enc$_i$ denotes the i-th layer of the pre-trained image transformer. We extract the embedding of [CLS] token as the output of the current layer. FFN is a Feed-Forward Network that aligns the dimensions of visual features to text features.

3.3.2 Multi-Modal Representation. Following previous works [16, 21, 26, 31], we adopted a self-attention transformer to accomplish multi-modal information interaction. Specifically, the input embedding of the multi-modal encoder is the sum of three components: (1) d_{V} or d_{T}, the visual or textual embedding initialized by a pre-trained image transformer or language transformer; (2) d_{m}, position embedding of the m-th token, where the beginning position ID of the tokens varies for different modalities; and (3) d_{t}, a segment embedding, which indicates whether the current token represents an image ($t = 1$) or text ($t = 0$). This approach ensures the integration of relevant information from different modalities while considering positional relationships and distinguishing between image and text tokens. We denote the input embedding of image as d_{V} and text as d_{T}, multi-modal transformer as f', then the multi-modal document representation can be denoted as:

$$\vec{X}'_{d;\phi} = f'_{\phi}([\mathrm{CLS}], d_{\mathrm{T}}, [\mathrm{SEP}], d_{\mathrm{V}}), \tag{6}$$

where $f'(\cdot)$ is the multi-modal page transformer, d_{T} and d_{V} are web page image and text representations. This approach culminates in the generation of more abundant representations by modalities complementing and reinforcing each other adequately in the multi-modal transformer.

The incorporation of visual information gives rise to a more abundant representation that encompasses the entirety of the web page, while the inclusion of block-level multi-modal information highlights vital blocks further and fosters a more comprehensive representation of the web page. In the subsequent section, we shall explicate the methodology for acquiring block-level information.

3.4 Incorporating Fine-Grained Block Representations

Owing to the intricacy of their components, web page screenshots offer a scant contribution to the precise semantics. To strengthen the semantics of specific content blocks and facilitate a more granular

representation of local context, a page segmentation method is used to partition the page into individual content blocks, encompassing paragraphs, headings, images, navigation menus, and other visually distinct blocks. A multi-modal transformer is utilized to generate block representations. By concatenating these multi-modal block representations, the multi-modal page transformer generates a more comprehensive representation of the entire web page.

3.4.1 Page Segmentation. Web pages consistently adhere to a structured format comprising distinct and meaningful blocks. These blocks possess a relatively independent semantic context, emphasizing the need to enhance comprehension of the essential blocks that may otherwise be overlooked within the coarse-grained web page modeling framework.

Leveraging the visual hierarchy and logical demarcation facilitated by the DOM tree, we use a DOM-based page segmentation algorithm. The algorithm recursively partitions the web page into blocks, utilizing both block text and screenshots as criteria. This combination enables the algorithm to determine the granularity of each block, facilitating the hierarchical division of the web page. Specifically, the overall process consists of two key steps: block tree construction and recursive segmentation. During block tree construction, our objective is to create a block tree wherein each node corresponds to a valid block on the web page. To accomplish this, we explore two options. First, we categorize DOM nodes as either block or inline elements based on their respective tags, selectively preserving nodes with block-level tags such as <p> and . Additionally, we eliminate nodes that pertain to the organizational and structural aspects of the HTML document but remain invisible on the rendered web page. These two approaches ensure that every node within the tree accurately maps to a valid block within the visual representation of the web page.

Once the block tree is constructed, we proceed with the recursive segmentation, following a top-down approach. The segmentation process begins by initializing a node pool that contains the nodes to be split. The root node of the block tree serves as the starting point for this process. In each round, a node is retrieved from the pool and assessed for its suitability for achieving the desired granularity. If the node does not meet the requirement, it is returned to the pool for further splitting. On the other hand, if the node satisfies the desired granularity, it is identified as a block in the final segmentation result. This approach enables a more detailed and granular understanding of the web page's structure and content. Further details regarding this segmentation will be presented in the subsequent section.

3.4.2 Multi-modal Block Representations. Upon segmentation of the web page into distinct blocks, we could encode each block to capture more granular multi-modal information. We leverage the same multi-modal transformer as introduced before to learn the representations of blocks. The block-level representations can be denoted as:

$$\begin{aligned} d_{\mathrm{B}} &= [d_{\mathrm{B}_1}, \cdots, d_{\mathrm{B}_k}], \\ d_{\mathrm{B}_i} &= f'_{\phi}([\mathrm{CLS}], d_{\mathrm{T}_{\mathrm{B}_i}}, [\mathrm{SEP}], d_{\mathrm{V}_{\mathrm{B}_i}}), \end{aligned} \tag{7}$$

where d_{B} are block-level multi-modal representations, $d_{\mathrm{T}_{\mathrm{B}_i}}$ and $d_{\mathrm{V}_{\mathrm{B}_i}}$ are text and image representations of the i-th block, respectively.

3.4.3 Block-Aware Multi-Modal Representation. To integrate fine-grained block representations into the overall document representation, a concatenation process is employed, combining all the multi-modal block representations with the page's text embeddings and image representations. Subsequently, they are jointly fed into the multi-modal page transformer. On one hand, the block representations enhance global semantic comprehension by incorporating local block semantics, thereby enriching the overall understanding of the page. On the other hand, the attention mechanism could lead to vital blocks being assigned higher attention and contributing more to the overall document representation.

Incorporating the fine-grained multi-modal block representations, the overall document representation can be denoted as:

$$\vec{X}''_{d;\phi} = f'_{\phi}([\mathrm{CLS}], d_{\mathrm{T}}, [\mathrm{SEP}], d_{\mathrm{V}}, [\mathrm{SEP}], d_{\mathrm{B}}), \tag{8}$$

where $f'(\cdot)$ is the multi-modal page transformer, d_{T} and d_{V} are web page image and text representations, and d_{B} is multi-modal representations of all blocks. In general, this model can effectively leverage the advantages of fine-grained block-level information. Compared to text-only models, our modal enables the retriever to create more abundant and comprehensive document representations. Consequently, it facilitates more effective matching and retrieval of relevant information.

3.5 Training Strategy

Following previous works of dense retrieval [13, 35, 38], we utilize prevalent contrastive learning strategies to train the dual-encoder. Concretely, contrastive loss can be calculated to optimize the dual-encoder for a given query q, a positive document d^+, and a set of negative documents $\mathcal{N}$ by maximizing the relevance score of the q and d^+ and minimizing that of the q and d^- in $\mathcal{N}$, i.e.

$$\mathcal{L}^q = -\log \frac{\exp(\mathcal{R}(q, d^+))}{\exp(\mathcal{R}(q, d^+)) + \sum_{d^- \in \mathcal{N}} \exp(\mathcal{R}(q, d^-))}, \tag{9}$$

where negative document set $\mathcal{N}$ could be documents in the same batch or top-ranked documents of a retrieve model [13, 35, 38]. The representations of d^+ and d^- can be obtained through Equation (8).

4 VILE DATASET CONSTRUCTION

The prevailing focus of existing datasets [3, 4, 24] pertaining to information retrieval predominantly revolves around textual content, thereby impeding the comprehensive exploration and advancement of models capable of effectively addressing the visual components inherent in web pages. In order to confront this quandary, scholars have proposed diverse methodologies and datasets [7, 34]. Nonetheless, these datasets suffer from either domain specificity [7] or insufficiency in size to support effective retrieval tasks [34]. Furthermore, these datasets disregard the abundant structural information inherent in web pages, solely providing visual snapshots, thus failing to meet the rigorous demands necessitated by the task of acquiring comprehensive representations for multi-modal documents.

To facilitate advanced research in fine-grained multi-modal information retrieval, we have constructed a novel dataset named VILE, encompassing both web page screenshots and block tree features, which together provide a comprehensive representation of the data. The VILE dataset is derived from the widely-adopted MS MARCO Document Ranking dataset [24], which is widely recognized as a

Table 1: Description of the VILE Dataset.

Description	File name	Size	#Records
Corpus Text	collections.tsv	1.9G	186,490
Corpus Screenshot	screenshot (.png)	412G	186,490
Corpus Block Tree	btree.tsv	90G	186,490
Query	queries.tsv	1.5M	39,194
Qrel Train	qrels.train.tsv	640k	35,661
Qrel Dev	qrels.dev.tsv	60K	3,533

benchmark for document retrieval tasks. The MS MARCO dataset consists of a substantial collection of 372 thousand queries and 3.2 million documents. Notably, all the queries in this dataset originate from real user queries submitted to the Bing search engine, with user identities being anonymized.

To construct the VILE dataset, we utilized a headless Firefox browser to visit the web pages indicated by the document URLs provided in the MS MARCO dataset. During this process, we rendered the pages, taking into account all the images and CSS styles present on the web page, and captured screenshots. The screenshots were standardized to have a fixed width, while the length remained unrestricted. Additionally, we traversed the DOM tree of each web page and recorded the two-dimensional positions of nodes within the overall web page screenshot. This approach enabled us to obtain the visual features of the web pages, with a minimum unit defined at the block level, ensuring a clear hierarchy of information. It is important to acknowledge that, due to the time-intensive nature of the rendering process and the presence of invalid URLs, we were able to assemble a corpus comprising 180 thousand documents. Among these documents, 40 thousand were identified as positive documents relevant to the query set. For a comprehensive understanding of the dataset, we refer interested readers to Table 1, which provides detailed information regarding its contents.

The characteristics of our dataset can be summarized as follows: (1). **Diverse Web Page Types**: By scraping web pages from real-world search engines, our dataset encompasses a wide range of web page types. This diversity is beneficial for training and evaluating models that aim to handle various web page layouts, designs, and content structures. (2). **Availability of Multiple Modals**: Our dataset includes not only the web page screenshots but also the corresponding block tree generated from the original DOM tree. This availability of multiple representations provides valuable information for parsing and understanding the structure of web pages. Researchers can utilize these components to develop models that leverage both textual and visual information effectively. (3). **Larger Corpus Size**: Compared to previous datasets [7, 34], our dataset offers a larger corpus containing over 180k web pages. The increased size is advantageous for training and evaluating retrieval models and allows researchers to explore more complex and sophisticated multi-modal models.

5 EXPERIMENTS

5.1 Baselines

We evaluate the performance of our approach by comparing it with three groups of methods for information retrieval:

(1). *Sparse Retrieval.* **BM25** [29] is a traditional but effective retrieval model based on probabilistic term counting. **Doc2query** [25] leverages a neural sequence-to-sequence model to expand documents with potential queries. **COIL** [10] uses contextualized representation to enhance exact term matching.

(2). *Visual Search.* **VIP** [7] incorporates visual information into Web search, which hybrids visual features together with LTR features to predict relevance score. **VITOR** [34] leverages a pre-trained image model to extract image features of web page screenshots. Besides, it extends screenshots by generating synthetic saliency heatmaps to better capture the attention information of web pages. Considering these models leverage LTR features as content features, we treat them as sparse retrieval models in the table.

(3). *Dense Retrieval.* This line of methods is based on dual-encoder architecture, which represents documents and queries with contextualized embeddings, and regards the dot products of query and document representations as relevance scores [30]. **RepBERT** [39] considers the documents in the same batch as negative samples. **DPR** [13] introduces the top 100 documents retrieved by BM25 as hard training negatives, which helps the model distinguish between positive documents and semantically similar documents. **ANCE** [35] retrieves hard negatives dynamically from the entire corpus using an asynchronously updated ANN index. Zhan et al. [38] proposes two training strategies respectively STAR and **ADORE**, the combination of which achieves best performance among these methods. Note that all of these models could be applied to our VILE framework, so as to have a thorough comparison. The results are listed in Table 2. Recently, some retrieval-oriented pre-trained language models win a lot of interest. Empowered by the flexibility of VILE, we could easily apply it to the pre-trained model. We list the results of **RetroMAE** [20] fine-tuned on the VILE dataset with the same settings as DPR.

5.2 Implementation Details

The experiments are carried out using an NVIDIA A40 (40GB) GPU. For all dense models, we employ the BERT-base [5] model as the encoder. Queries are truncated to a maximum of 32 tokens, while documents are truncated to a maximum of 256 tokens. For the Visual-Enc component of VILE, we employ a vision transformer that has been pre-trained by a multi-modal framework CLIP [28]. Furthermore, in our ablation study, we also investigate the use of a vision transformer pre-trained using the VIT framework [6] to evaluate the impact of different pre-training approaches for image encoder. To preserve visual features at different semantic levels, we extract the [CLS] embedding from each transformer layer. Subsequently, we employ a four-layer FFN to project the image features to a dimension of 768, which we then concatenate with the text features. Inspired by previous multi-modal works [16, 21], we initialize the weights of multi-modal transformer with BERT-base.

To replicate the baselines, we utilize open-source repositories [18, 22] for sparse retrieval models. For all dense retrieval methods, we employ a dual-tower encoder initialized with BERT-base. To ensure a fair comparison, we optimize all dense retrievers using the Lamb optimizer [37], with a batch size of 32 and one hard negative sample. For text-only models, we utilize list-wise cross-entropy loss and set the learning rate to $3e$-5 choosing from $1e$-5 to $5e$-4. For

Table 2: Results of all models on VILE Dataset. We show the improvement of our model compared with corresponding text-only models. "T" and "V" denotes text information and visual information, respectively. "†" denotes the result is significantly better than corresponding text-only models in t-test with $p < 0.05$ level.

Model Type	Model	Type	MRR@10		MRR@100		RECALL@1		RECALL@10		RECALL@100	
Sparse	BM25	T	0.3158		0.3256		0.2327		0.5078		0.7499	
	Doc2query	T	0.3372		0.3464		0.2502		0.5389		0.7662	
	COIL	T	0.5121		0.5176		0.4142		0.7135		0.8334	
	VIP	T & V	0.3961		0.4053		0.3133		0.5800		0.7698	
	VITOR	T & V	0.4077		0.4166		0.3247		0.5893		0.7753	
Dense	RepBERT	T	0.4994	-	0.5051	-	0.4104	-	0.6810	-	0.8092	-
	RepBERT+**VILE**	T & V	0.5227†	+4.67%	0.5283†	+4.59%	0.4301†	+4.80%	0.7144†	+4.90%	0.8287†	+2.41%
	DPR	T	0.5082	-	0.5134	-	0.4166	-	0.6946	-	0.8235	-
	DPR+**VILE**	T & V	0.5252†	+3.35%	0.5306†	+3.35%	0.4370†	+4.90%	0.7143†	+2.84%	0.8291	+0.68%
	ANCE	T	0.5212	-	0.5260	-	0.4291	-	0.7093	-	0.8272	-
	ANCE+**VILE**	T & V	0.5346†	+2.57%	0.5393†	+2.53%	0.4452†	+3.75%	0.7158†	+0.93%	0.8316	+0.53%
	ADORE	T	0.5335	-	0.5383	-	0.4458	-	0.7116	-	0.8294	-
	ADORE+**VILE**	T & V	0.5430†	+1.78%	0.5475†	+1.71%	0.4534†	+1.70%	0.7206†	+1.26%	0.8332	+0.46%
	RetroMAE	T	0.5338	-	0.5384	-	0.4390	-	0.7201	-	0.8296	-
	RetroMAE+**VILE**	T & V	**0.5668**†	+6.18%	**0.5716**†	+6.17%	**0.4750**†	+8.20%	**0.7424**†	+3.10%	**0.8472**†	+2.12%

VILE models, we utilize the same loss and set the learning rate to $1e$-4 choosing from $1e$-5 to $5e$-4. ADORE and ADORE+VILE use the same settings as the original paper, with the AdamW [15] optimizer, a learning rate of $5e$-6, and a batch size of 32. RetroMAE and RetroMAE+VILE are fine-tuned in the VILE dataset with the same setting as DPR and DPR+VILE. The Faiss library [12] is employed for efficient similarity search and calculating inner products.

5.3 Overall Results

Experimental results on the VILE dataset are shown in Table 2. We provide a comprehensive analysis by reporting various evaluation metrics to facilitate a thorough comparison of the retrieval results. These metrics include MRR@10, MRR@100, RECALL@1, RECALL@10, and RECALL@100. Some observations are summarized as follows.

(1) Among all the dense retrievers, **our visual enhanced models** (denoted as "**+VILE**") **significantly outperform their corresponding text-only baselines**, especially on top-ranking performance. Compared with the best baseline models, our models have significant improvements with paired t-tests at $p < 0.05$ level on most metrics. For example, it outperforms RepBERT by 4.67% and 4.90% in terms of MRR@10 and Recall@10, respectively. Among all the dense retrieval methods initialized by BERT, the combination of ADORE and VILE achieves the best performance, with improvements of 1.78% and 1.26% on MRR@10 and RECALL@10, respectively. These results highlight that VILE can serve as a powerful document representation framework by incorporating visual information into the document representation, consistently enhancing retrieval performance across various negative sampling approaches.

(2) VILE significantly outperforms previous visual search methods. VIP and VITOR incorporate visual information into LTR content features to enhance retrieval, surpassing text-only sparse retrieval baselines BM25 and Doc2query. VILE models outperform them with an MRR@10 increase of over 30%. This highlights the effectiveness of leveraging fine-grained interactions between visual features and text features, as it allows visual features to unleash their full potential in enhancing document representations and ultimately improving retrieval performance.

(3) VILE has a strong adaptation to retrieval-oriented pre-trained models. Retrieval-oriented pre-trained models can improve the effectiveness of retrievers by enhancing the representation of [CLS] embedding through pre-training. RetroMAE+VILE outperforms text-only RetroMAE by 6.18% on MRR@10 and 8.20% on Recall@1. This demonstrates that VILE exhibits a robust adaptation capability to various types of models, including retrieval-oriented pre-trained models.

In summary, the experimental results validate the effectiveness of VILE in various negative sampling methods and pre-trained models, showcasing its potential in enhancing document retrieval. Moreover, the advantage of VILE lies in the offline calculation of document representations, which can be stored in an inverted index without incurring additional storage or inference time costs.

5.4 Further Analysis

In order to delve deeper into the intricacies of VILE, we have conducted additional experiments aimed at analyzing the influence of visual information and block-level information. **In consideration of training efficiency, we have selected RepBERT as the text-only model and RepBERT+VILE as the multi-modal model, with the latter abbreviated as VILE.**

In this section, we first assess how visual and block-level information affect model performance. Then we examine image encoders' retrieval performance using different pre-training methods and block granularity. We also study VILE's best retrieval scenarios in detail. We visualize our multi-modal page transformer's block attention weights to determine block importance. Finally, we demonstrate VILE's efficacy and applicability in real-world retrieval tasks by conducting case studies. These additional experiments help us understand VILE's strengths and distinctive traits, making it a useful model for web page retrieval.

Table 3: Results with different modal information used.

Model	MRR@10		RECALL@10		RECALL@100	
VILE	0.5227	-	0.7144	-	0.8287	-
w/o. Visual	0.5092	-2.58%	0.6923	-3.09%	0.8132	-1.87%
w/o. Block	0.5060	-3.19%	0.6897	-3.46%	0.8125	-1.95%
w. $\text{Block}_{\text{Avg}}$	0.5094	-2.54%	0.6987	-2.20%	0.8128	-1.92%
w. $\text{Block}_{\text{Max}}$	0.5113	-2.18%	0.7011	-1.86%	0.8133	-1.86%

Table 4: Performance of different visual encoders. RepBERT+VILE is referred to as VILE. "†" denotes the result is significantly better than baseline with $p < 0.05$ level.

Model	MRR@10		RECALL@10		RECALL@100	
RepBERT	0.4994	-	0.6810	-	0.8092	-
$\text{VILE}_{\text{CLIP}}$	$0.5227^{\dagger}$	+4.67%	$0.7144^{\dagger}$	+4.90%	$0.8287^{\dagger}$	+2.41%
VILE_{VIT}	$0.5180^{\dagger}$	+3.72%	$0.7098^{\dagger}$	+4.23%	$0.8258^{\dagger}$	+2.05%

5.4.1 Ablation Study. In this paper, VILE mainly uses two components to incorporate visual information into document representation of retrieval. Firstly, the multi-modal page transformer learns a more abundant representation of documents by incorporating visual information. Secondly, VILE segments the entire page into blocks, and aggregates the fine-grained block-level information to facilitate a more comprehensive representation. To eliminate the effect of visual information, we mask the visual features in the block transformer and page transformer, which will make sure the input sequence length of the variant is the same and result in a fair comparison, denoted as *w/o.* Visual. To eliminate the effect of block-level information, we remove the block encoder, and only leverage the multi-modal page transformer to learn a document representation, denoted as *w/o.* Block. Besides, to verify the effectiveness of the hierarchical fusion architecture, we explore two score aggregation methods. We first fed the whole page and all blocks into the multi-modal transformer, calculate the similarity scores, and obtain page score S1 and block score list. Then two aggregation methods are used to process the block score list and get score S2, respectively average aggregation (denoted as *w.* $\text{Block}_{\text{Avg}}$) and max aggregation (denoted as *w.* $\text{Block}_{\text{Max}}$). Final document similarity is a weighted sum of S1 and S2.

The experiment results in Table 3 demonstrate the impact of removing components on VILE's performance across various metrics. The findings are as follows: (1) Removing each component negatively affects performance. Without visual information, there is a significant loss of 3.09% on Recall@10 and 2.58% on MRR@10, indicating the effectiveness of visual information. Removing block information results in a loss of 3.46% on Recall@10 and 3.19% on MRR@10, highlighting the importance of page segmentation. (2) Block information improves retrieval. Both *w.* $\text{Block}_{\text{Max}}$ and *w.* $\text{Block}_{\text{Avg}}$ outperform *w/o.* Block, with *w.* $\text{Block}_{\text{Max}}$ showing a 3.19% improvement on Recall@10. (3) Hierarchical fusion enhances retrieval. Both aggregations are less effective than VILE, with *w.* $\text{Block}_{\text{Avg}}$ experiencing a loss of 2.54% on MRR@10 and 2.20% on Recall@10. *w.* $\text{Block}_{\text{Max}}$ performs better but still experiences a loss of 2.18% on MRR@10 and 1.86% on Recall@10. These findings underscore the importance of visual information and detailed block segmentation in improving VILE's retrieval performance. By incorporating these components, VILE effectively leverages visual features, captures detailed information, and enhances retrieval capabilities.

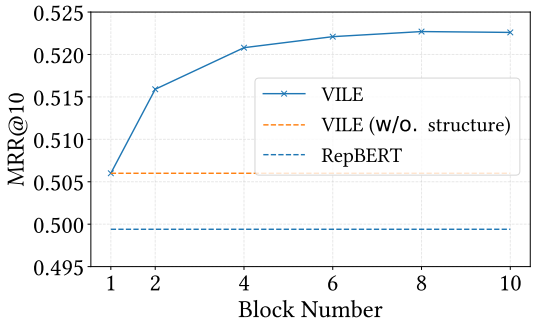

Figure 3: The effect of block granularity. Block number can be controlled by the segmentation method in Section 3.4.1.

5.4.2 The Effect of Image Encoder. Additionally, we conducted experiments to evaluate the influence of different visual encoders. As illustrated in Table 4, VILE initialized by both CLIP and VIT yields significant enhancements compared to RepBERT. Between these two models, CLIP is initialized through a multi-modal pre-training task, whereas VIT is pre-trained solely on image-related tasks. The findings indicate that VILE_{VIT} achieves slightly inferior performance to $\text{VILE}_{\text{CLIP}}$, with a decrease of 0.90% in MRR@10 and 0.64% in Recall@1. This implies that the visual features acquired through multi-modal pre-training are more suitable for our specific task.

5.4.3 The Effect of Block Granularity. Block granularity, the level of detail or size of web page blocks, is a crucial hyper-parameter in block-enhanced visual representation. Smaller blocks enhance local block information transmission but increase computational cost due to more blocks in web page segmentation. In Section 3.4.1, we describe a method to control block granularity by defining the partition criterion during recursive segmentation. By adjusting the criterion, we control block size and number. Larger text and screenshot sizes result in fewer blocks. We select the top k largest blocks as the final representation for each web page, specifically $\{1, 2, 4, 6, 8, 10\}$. Setting k to 1 removes block information from VILE.

We analyze the relationship between block number and retrieval effectiveness on MRR@10 (Figure 3). Intuitively, increasing block granularity should improve retrieval performance. However, we find that after reaching 8 blocks, the performance improvement becomes less significant and may decline. This suggests that too many blocks can disrupt semantic coherence within the web page. To balance effectiveness and efficiency, we set 8 as the maximum number of blocks for our experiments.

5.4.4 Block Attention Visualization. In this section, we visualize the page segmentation result and block attention weight. Attention

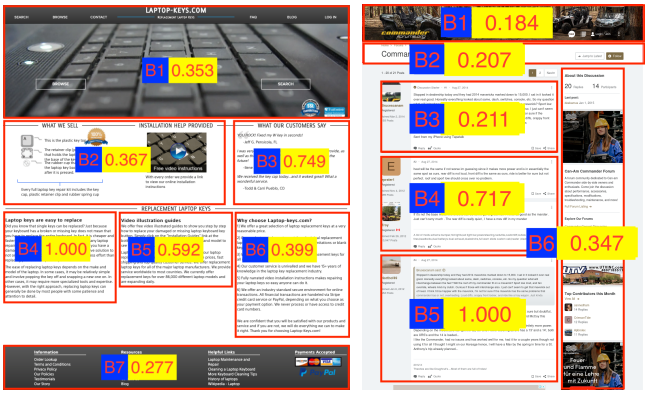

Figure 4: The segmentation results and block attention weights of documents D564733 and D17021.

weights are extracted from the [CLS] token position of the multi-modal page transformer, reflecting block contributions. To facilitate comparison, attention weights are exponentially amplified by a factor of 1,000 and normalized. Figure 4 shows the visualization of normalized attention weights. In the left web page, seven blocks are identified, with B4 having the highest weight. Block B4's content, titled "Laptop keys are easy to replace", is highly relevant to the web page's title, "replace laptop keys". However, due to length limitations, models may overlook B4, leading to inaccurate relevance assessments. Header block B1 and footer block B7 have lower weights due to their shorter text. However, B1 is still more relevant, featuring a keyboard figure and the title "Laptop-keys.com–replace laptop keys". Similarly, the right web page has six blocks, with B5 having the highest weight. This highlights the importance of block segmentation and VILE's ability to distinguish relevant blocks from irrelevant ones.

5.4.5 Case Study. To compare the retrieval results of VILE and RepBERT, we present a case query with their rankings in Table 5. The query is "how to slow cook beef ribs in the oven". The ground-truth document (D132881) titled "Slow Cooker Barbequed Beef Ribs" provides step-by-step cooking instructions with subtitles and vivid figure demonstrations. In contrast, a similar document (D199716) titled "How to cook fall-off-the-bone beef spare ribs in the oven or on the stovetop" has less structured text without illustrations. Considering the visual information, D132881 is more relevant to the query, featuring well-structured text and relevant illustrations. This showcases the benefits of vision-enhanced representation, enabling VILE to accurately retrieve the golden document as the top-ranked result. RepBERT, which ignores structural information, fails to differentiate between these two documents.

6 CONCLUSION AND FUTURE WORK

In conclusion, we propose VILE, a new dense retriever that addresses the limitations of existing approaches by incorporating visual features into document representations for web search. While previous models focus primarily on textual content, disregarding important visual features such as style, images, and layout, VILE

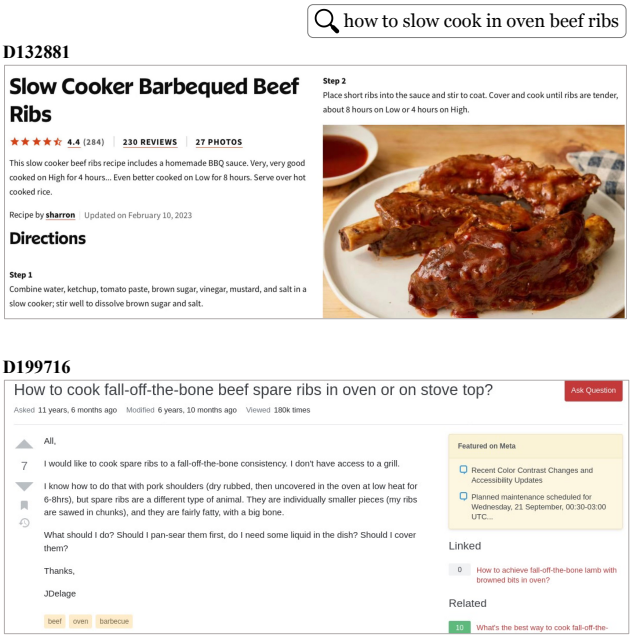

Figure 5: A case study: D132881 (ground-truth) is ranked 1st and 8th by VILE and RepBERT; D199716 is a similar document, which is ranked 2nd and 1st by VILE and RepBERT.

takes a multi-modal approach to capture both textual and visual information comprehensively. Furthermore, by segmenting web pages into blocks and creating multi-modal representations for each block, VILE effectively captures the fine-grained characteristics of different content regions. These block representations are then integrated to form a comprehensive multi-modal representation of the entire web page. Experimental results on a newly constructed multi-modal document retrieval dataset demonstrate significant performance improvements over existing models, indicating the effectiveness of VILE in capturing and understanding diverse content regions. Moving forward, we aim to explore large-scale multi-modal pre-trained models to further enhance the quality of document representations.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work was supported by the National Natural Science Foundation of China No. 62272467 and No. 61832017, Beijing Outstanding Young Scientist Program No. BJJWZYJH012019100020098, the fund for building world-class universities (disciplines) of Renmin University of China, Public Computing Cloud, Renmin University of China, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. The work was partially done at Beijing Key Laboratory of Big Data Management and Analysis Methods, and Key Laboratory of Data Engineering and Knowledge Engineering, MOE.

REFERENCES

[1] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932* (2020).

[2] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[3] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview.* Technical Report. MICHIGAN UNIV ANN ARBOR.

[4] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Liang Pang, and Xueqi Cheng. 2017. Learning visual features from snapshots for web search. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 247–256.

[8] Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253* (2021).

[9] Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540* (2021).

[10] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186* (2021).

[11] Rebecca A Grier. 2004. *Visual attention and web design.* University of Cincinnati.

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[13] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[14] Andy J King, Allison J Lazard, and Shawna R White. 2020. The influence of visual complexity on initial user impressions: Testing the persuasive model of web design. *Behaviour & Information Technology* 39, 5 (2020), 497–510.

[15] Diederik P Kingma, J Adam Ba, and J Adam. 2020. A method for stochastic optimization. arXiv 2014. *arXiv preprint arXiv:1412.6980* 106 (2020).

[16] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[17] Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More robust dense retrieval with contrastive dual learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 287–296.

[18] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2356–2362.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Zheng Liu and Yingxia Shao. 2022. Retromae: Pre-training retrieval-oriented transformers via masked auto-encoder. *arXiv preprint arXiv:2205.12035* (2022).

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[22] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation inInformation Retrieval using PyTerrier. In *Proceedings of ICTIR 2020.*

[23] Mandar Mitra and BB Chaudhuri. 2000. Information retrieval from documents: A survey. *Information retrieval* 2 (2000), 141–163.

[24] Tri Nguyen, Mir Rosenberg, Xia Song, et al. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *NIPS 2016.*

[25] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375* (2019).

[26] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966* (2020).

[27] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2010. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval* 13 (2010), 346–374.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[29] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[30] Fumin Shen, Wei Liu, Shaoting Zhang, Yang Yang, and Heng Tao Shen. 2015. Learning binary codes for maximum inner product search. In *Proceedings of the IEEE International Conference on Computer Vision*. 4148–4156.

[31] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

[32] Lisbeth Thorlacius. 2010. Visual communication in web design–analyzing visual communication in web design. *International handbook of internet research* (2010), 455–476.

[33] Mohamed Trabelsi, Zhiyu Chen, Brian D Davison, and Jeff Heflin. 2021. Neural ranking models for document retrieval. *Information Retrieval Journal* 24 (2021), 400–444.

[34] Bram van den Akker, Ilya Markov, and Maarten de Rijke. 2019. ViTOR: learning to rank webpages based on visual features. In *The world wide web conference*. 3279–3285.

[35] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[36] Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of BERT for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972* (2019).

[37] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962* (2019).

[38] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1503–1512.

[39] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498* (2020).

[40] Junqi Zhang, Yiqun Liu, Shaoping Ma, and Qi Tian. 2018. Relevance estimation with multiple information sources on search engine result pages. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 627–636.