UTRECHT UNIVERSITY

DIGITAL RISK SOLUTIONS DELOITTE NL

MASTER THESIS

A Multi-Modal Approach to Open Domain Question Answering: Dense Retrieval of Regions-of-Interest

Author:
Massimiliano Garzoni di Adorgnano

Academic supervisor:
Dr. Albert Gatt

Company supervisor:
Niels van der Heijden

Second reader:
Dr. Heysem Kaya

*A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Artificial Intelligence
in the Department of Information and Computing Sciences
at Utrecht University*

October 17, 2023

UTRECHT UNIVERSITY

Abstract

Faculty of Science
Department of Information and Computing Sciences

Master of Science in Artificial Intelligence

A Multi-Modal Approach to Open Domain Question Answering: Dense Retrieval of Regions-of-Interest

by Massimiliano Garzoni di Adorgnano

This master thesis proposes a multi-modal retrieval system for open-domain question answering, building upon dense passage retrievers and incorporating multi-modal information to retrieve candidate regions of interest (ROIs) from document images given a user query. Our main research goal was to investigate the efficacy of dense representations of questions and multi-modal contexts in retrieving relevant content, and evaluating the impact of multi-modal information compared to uni-modal baselines. To this end, the study leverages the VisualMRC dataset which offers annotations for visual components, particularly ROIs such as titles or graphs, to facilitate efficient content retrieval. The proposed methodology involves pre-processing the multi-modal ROIs, employing a bi-encoder setup to encode the question and ROIs separately, and use such encodings to calculate similarity in their shared multi-dimensional embedding space. The training objective is achieved through contrastive learning by passing to the model a question, along with one positive and k negative contexts, and minimizing the loss function by reducing the negative log likelihood associated with the positive ROI. We evaluate our trained models on three different modality scenarios, text-only, vision-only, and multi-modal, and we evaluate their retrieval performance on standard metrics such as Normalized Cumulative Discounted Gain @ k, Mean Reciprocal Rank @ k, and Recall @ k. The results reveal the benefits of both vision-only and multi-modal approaches over text-only, while also highlighting challenges related to the number of negative ROIs. Our results support the first hypothesis but raise questions about the second, suggesting that the inclusion of layout information may not always improve retrieval performance. The strengths of our approach include efficient ROI retrieval and dataset adaptability, while limitations involve dataset variability and encoding techniques. In light of this, we suggest several avenues for future work such as exploring new datasets, incorporating hard negatives in contrastive learning, and refining ROI dissimilarity. Additionally, we speculate that integrating keyword matching and retrieval-augmented generation approaches could enhance the retrieval pipeline. Overall, the present thesis hopes to advance research in multi-modal retrieval models, emphasizing the importance of visual and textual context for open-domain question answering.

Acknowledgements

Since no great things are done alone, I would like to thank a few people who have been very important for me and for this challenging project during these months.

First and foremost, I want to thank my family for the support throughout the years: you gave me the opportunity to study and always encouraged me to pursue my interests, always believing in me. Papá, Mamma, Cate, Bibi e Fede: grazie di tutto.

I thank and am grateful for my supervisor Dr. Gatt who has provided very actionable, useful and critical feedback during the various stages of the thesis project. Similarly, I thank Dr. Kaya for additional support and feedback on the thesis proposal and arrangement of the project. I learnt a lot from you both.

Importantly, I want to thank Niels and Polina from the DRS team at Deloitte, who supported me and provided feedback on a daily basis. I learnt a lot from you and I am grateful to have had the chance to work with you, and the rest of the DocQMiner team.

I want to thank my dear friend and housemate François for sharing the thesis struggle with me - we got this! I also would like to thank my university and work colleague Nacho, who shared programming and technical knowledge, provided useful feedback and cheered me up when I needed it. Not to forget my closet friends back home - Fedoc, Jack and Andre - who were always close to me even when physically far.

Lastly, and certainly not least, I thank you, Elena: your constant love and support always swept away the dark clouds and brought back light and motivation, even in the darkest times.

Finally, I would like to thank myself for achieving such challenging project and being open to learn from all the setbacks to improve for the future.

Contents

List of Abbreviations

QA	**Q**uestion **A**nswering
NLP	**N**atural **L**anguage **P**rocessing
IR	**I**nformation **R**etrieval
MRC	**M**achine **R**eading **C**omprehension
DPR	**D**ense **P**assage **R**etrieval
VDU	**V**isual **D**ocument **U**nderstanding
VQA	**V**isual **Q**uestion **A**nswering
ROI(s)	**R**egion(s) **o**f **I**nterest
OCR	**O**ptical **C**haracter **R**ecognition

Dedicated to my parents and grandparents, who taught me passion, curiosity and perseverance.

Chapter 1

Introduction and motivations

The ability to ask and answer questions is a fundamental aspect of human intelligence, as it allows us to explore the world, to learn new things, and to make sense of our experiences. This capacity to inquire is perhaps what separates us from other animals, and it has been instrumental for our development as a species. Perhaps more important for that, however, is the ability to efficiently retrieve relevant information from our memory or external sources in order to provide meaningful answers. In the real world, this is done in a multi-modal fashion, and thus the incorporation of multi-modal information into a question answering framework from documents entails utilizing various sensory inputs, particularly images and their layouts, to enhance the performance of information retrieval. When combined with textual data, these visual cues provide a richer contextual background, enabling question answering models to better understand the query. Integrating such information to represent the input document image enables the model to grasp not only the meaning of the text but also insights from visual elements. This integration enhances information retrieval but also allows for more nuanced and accurate answers, as we expect it to surpass the constraints of text-only retrieval methods.

If we are to recreate this general capacity in computational systems, we must first appreciate its significance and complexity. Human language is inherently ambiguous and context-dependent, and understanding it requires more than just matching keywords. In fact, to answer a question, a system must be able to grasp the nuances of meaning and draw on a wide range of knowledge sources. In addition, it must be able to reason, infer, and generate new insights, much like a human would.

1.1 Aim of the project

In this work we propose a retrieval system which expands upon dense passage retrievers (specifically DPR [13]) and incorporates multi-modal information to select candidate regions of interest (ROI(s)) present in document images (i.e., titles, paragraphs, graphs, tables) from which the answer to an input question may be extracted or generated by reasoning upon that region in the document.

1.1.1 Research questions

For the purpose of this project, we define the following key research questions which will be tackled:

- **RQ1**: Can we devise a method that uses dense representations of questions and multi-modal contexts (i.e., regions or interest from a document image such as titles and graphs) to efficiently retrieve the contexts that are more likely to contain the answer to a given question, i.e. relevant to the question?

- **RQ2**: Does multi-modal information improve retrieval performance when compared to uni-modal (text-only and image-only) baselines?

1.1.2 Hypotheses

- **H1**: By expanding on the methodology devised in DPR [13], we expect to be able to treat multi-modal contexts as passages and use them to inform the reasoning for answering the question.
- **H2**: We expect multi-modal information (joint representation of text, image and layout) to improve retrieval performance over uni-modal approaches.

1.2 Motivations

Information retrieval is a fundamental task in the digital age, but traditional text-based approaches have limitations in providing comprehensive and contextually relevant answers to user queries. This thesis aims to address the shortcomings of text-based information retrieval by exploring the integration of multi-modal data sources to enhance the retrieval accuracy and richness of responses.

1.2.1 Scientific motivation

Research on the comparison of uni-modal and multi-modal encoding mechanisms for information retrieval from document images is of great significance for academia and the scientific community. Within academia, the outcomes of this study can offer valuable insights into the development of neural network models for document visual information retrieval, leading to improved accuracy and robustness. Specifically, emphasizing the significance of visual layout and structure alongside textual context expands the current research efforts in this domain. For the scientific community, advancements in this field have the potential to transform information accessibility and utilization, making it more widely available and easily accessible to various stakeholders. Furthermore, the findings from this research may have practical applications in other disciplines like medical imaging, law, or finance. This research aims to provide further understanding of the effectiveness of unified multi-modal approaches and inspire further advancements in this direction.

1.2.2 Motivation for Deloitte's Digital Risk Solutions and DocQMiner

The present research work is relevant to a range of industry applications, particularly in the area of document information extraction and understanding. For example, such systems can be devised for both legal and financial applications to extract and comprehend information from legal documents and financial reports, respectively. By leveraging uni-modal and multi-modal methods, these systems can accurately identify key terms, clauses, financial performance metrics, and industry trends from large volumes of documents. This can increase the efficiency of legal research and document reviews, as well as financial analysis and decision-making, freeing up time for experts to focus on value-added tasks. Both of these are use cases for the DocQMiner product offered by Deloitte's Digital Risk Solutions, from which this research work is sponsored.

Chapter 2

Background and Related Work

2.1 Background

Before diving into the methodology and experimental aspects of this study, it is important to establish the groundwork by introducing relevant theoretical foundations and lines of research that have inspired the current work. This serves as a fundamental framework for comprehending and contextualizing the subsequent chapters.

2.1.1 Information Retrieval

Information Retrieval (IR) is a field dedicated to the organization and retrieval of information from vast collections of documents. It involves developing methodologies and algorithms to effectively and efficiently retrieve relevant documents in response to user queries. In the context of retrieval from documents, IR focuses on understanding user questions and retrieving specific pieces of information from the document collection that directly address those questions. Various methodologies are employed to solve this task, including keyword matching, vector space models [29], probabilistic models [5], and more recently, deep learning-based approaches using transformer models. The latter, specially those based on transformers, have gained prominence due to their ability to handle intricate inter-modal relationships and capture semantically rich contextual information [20]. Sparse vector-based techniques, such as Latent Dirichlet Allocation (LDA), offer a scalable means of capturing cross-modal associations by projecting data into shared vector spaces [29]. On the other hand, dense-vector or deep learning-based methodologies, take advantage of large-scale datasets to learn complex patterns and generate joint representations, enhancing the capacity to handle diverse queries and data formats.

The scope of IR extends to encompass diverse media types, paving the way for multi-modal IR, which leverages multiple types of data such as text, images, audio, and videos to enhance the retrieval process. The main assumption behind multi-modal IR is that different modalities can contain complementary information, thereby potentially leading to more comprehensive and accurate search results. This approach is particularly useful when queries may be ambiguous or difficult to express solely through text, thus considering various modalities can offer a more holistic understanding of user intent.

2.1.2 Question Answering

Question answering (QA) is a multidisciplinary field which aims to develop computer systems capable of providing relevant answers to natural language questions posed by users. Due to the complexity of QA, it has been studied under various tasks and circumstances, leading to sub-fields such as Machine Reading Comprehension (MRC) and Document Visual Question Answering (Document VQA), as well as the distinction between extractive and generative QA systems, among others.

Early research in QA focused on creating closed-domain systems that relied on a core knowledge database (KB) predefined by developers. These systems utilized the KB for question interpretation and answer retrieval. Examples of such systems include BASEBALL [10], LUNAR [36], and MURAX [17]. In the early 2000s, the concept of utilizing the World Wide Web as a source of information emerged. Systems like MULDER [18] automatically responded to open-domain questions by leveraging search engines. This shift from specific databases to web-based information allowed systems to consider a greater amount of contextual information, resulting in improved accuracy of answers.

MRC focuses on enabling machines to comprehend natural language from one or more human-generated text sources. In the context of QA, MRC involves processing a given text passage and answering questions related to it. Deep learning models, such as neural networks, are trained to learn the relationship between the input text and the answer to a question. Recent years have seen significant research in MRC, facilitated by large-scale datasets like the Stanford Question Answering Dataset (SQuAD) [27] and benchmark tasks such as the Microsoft Machine Reading Comprehension (MS MARCO) challenge [2]. These resources have paved the way for sophisticated MRC models applicable to diverse domains, including news articles, scientific papers, and legal documents. The VisualMRC dataset, which will be discussed in Section 2.2.3, represents a significant effort in this area.

Another important distinction in QA lies between extractive (EQA) and generative (GQA) question answering tasks. In EQA, the system extracts the answer from a given text by identifying a relevant span that matches the question. In contrast, GQA involves generating a new text that answers the question. Extractive approaches are more common due to their ability to directly retrieve answers from the context. On the other hand, generative QA is more challenging as it requires the system to comprehend the question and generate a coherent and relevant response in natural language.

2.1.3 Open-Domain QA

Although numerous techniques have been developed to enhance QA system performance, notable achievements were primarily limited to closed-domain scenarios (as discussed in Section 2.1.2) [31]. However, the field has progressed to address the demand for QA systems capable of handling questions across various domains. This gave rise to open-domain QA, where systems aim to answer natural language questions on a wide range of topics without relying on a predefined knowledge base. Instead, they utilize a large, unstructured corpus of text documents as context. This entails understanding the user's question and retrieving relevant information from sources such as web pages, articles, and books.

To tackle open-domain QA, researchers have increasingly turned to statistical learning approaches for extracting potential answers from vast collections of unstructured documents. As a result, the effectiveness of QA systems has significantly improved for open-domain questions. These systems retrieve the most probable response by matching and retrieving information from an extensive knowledge base, including websites and research papers.

In the context of open-domain question answering, MRC algorithms are employed to automatically *read* and *understand* large amounts of text, such as web pages or collections of legal documents, and provide answers to user questions. This type of QA, often referred to as reading comprehension across multiple documents [34], has the potential to revolutionize information access and processing by enabling faster and more efficient handling of complex questions.

2.1.4 (Computational) Multi-Modality

Transformer models have revolutionized the field of natural language processing (NLP) by demonstrating exceptional performance on various tasks. However, these models have traditionally focused on single modalities for perception and understanding tasks, neglecting the potential benefits of multi-modality [24, 31]. In contrast, biological systems excel at perceiving and interacting with the environment through various sensory inputs, combining information from distinct modalities. To bridge this gap and enable computer systems to leverage multi-modal information, researchers have developed transformer-based models that incorporate multiple sensory inputs. Two notable approaches are dual encoders and fusion encoders, which differ in how they integrate modalities within the model architecture.

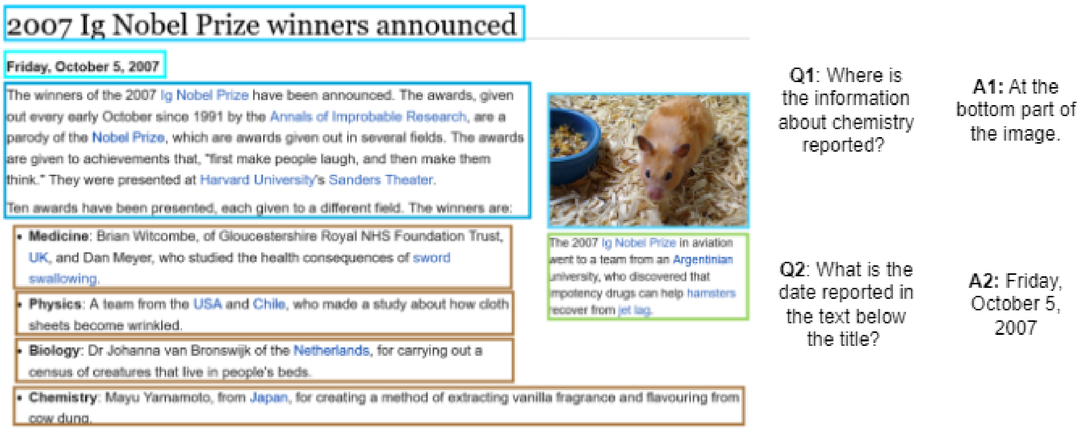
Dual encoders, as the name suggests, keep the modalities separate throughout the model layers and only mix the modality information at the final layer. These models maintain separate encoder paths for different modalities and merge them at the last layer to generate meaningful outputs. This approach allows each modality to undergo separate processing, preserving their individual characteristics and capturing their unique features. Examples of dual encoder, multi-modal approaches include the Vision Transformer (ViT) [7], ViLT [14] and ViLBERT [21].

On the other hand, fusion encoders aim to fuse the modalities at an earlier stage of the model architecture. These models directly combine the representations of different modalities from the input stage and jointly process them throughout the model layers. By integrating the modalities from the beginning, fusion encoders can capture complex interactions and dependencies between modalities more effectively. This approach enables better exploitation of the complementary information provided by each modality. The work in [23] present notable examples of fusion-based approaches.

In terms of pre-training methods, both dual encoders and fusion encoders can benefit from pre-training on large-scale multi-modal datasets. The pre-training process involves exposing the models to a vast amount of multi-modal data, allowing them to learn rich representations of both textual and visual information. This pre-training helps the models develop a comprehensive understanding of complex phenomena by jointly learning from multiple modalities. We expand more on this matter in Section 2.2.2. When comparing the two approaches, dual encoders have the advantage of preserving the unique characteristics of individual modalities, enabling finer-grained control over each modality's representation [8]. On the other hand, fusion encoders offer the benefit of capturing intricate inter-dependencies between modalities from the beginning of the model, potentially leading to better performance in tasks that require strong cross-modal interactions [8].

Consider for example the QA pair in Figure 2.1. These questions refer to physical regions of the image, so they can be more precisely answered by including visual information rather than just textual context. By incorporating the visual modality, the system ought to better comprehend the query and provide a more comprehensive and precise answer.

An interesting effort in this direction is the work proposed in the ManyModalQA challenge [11], where they present a dataset collected by scraping Wikipedia and crowd-sourcing question-answer pairs. The intriguing aspect of this dataset is that the questions are intentionally ambiguous, making it difficult to determine the modality containing the answer based solely on the question. The authors define a selector



2007 Ig Nobel Prize winners announced

Friday, October 5, 2007

The winners of the 2007 Ig Nobel Prize have been announced. The awards, given out every early October since 1991 by the Annals of Improbable Research, are a parody of the Nobel Prize, which are awards given out in several fields. The awards are given to achievements that, "first make people laugh, and then make them think." They were presented at Harvard University's Sanders Theater.

Ten awards have been presented, each given to a different field. The winners are:

- **Medicine**: Brian Witcombe, of Gloucestershire Royal NHS Foundation Trust, UK, and Dan Meyer, who studied the health consequences of sword swallowing.
- **Physics**: A team from the USA and Chile, who made a study about how cloth sheets become wrinkled.
- **Biology**: Dr Johanna van Bronswijk of the Netherlands, for carrying out a census of creatures that live in people's beds.
- **Chemistry**: Mayu Yamamoto, from Japan, for creating a method of extracting vanilla fragrance and flavouring from cow dung.

The 2007 Ig Nobel Prize in aviation went to a team from an Argentinian university, who discovered that impotency drugs can help hamsters recover from jet lag.

Q1: Where is the information about chemistry reported?

A1: At the bottom part of the image.

Q2: What is the date reported in the text below the title?

A2: Friday, October 5, 2007

FIGURE 2.1: An example document image and its corresponding Regions-Of-Interest. Image retrieved from [33] and QA pairs provided by the author.

network is constructed to predict the relevant modality for the answer, revealing that the dataset's questions are more ambiguous than existing datasets. An important concept at the heart of the present work is modality disambiguation, which refers to the process of determining the most appropriate or relevant modality (such as text, images, tables, audio, etc.) for extracting information required to answer a question or solve a problem, thus aiming to identify which specific modality or combination of modalities holds the answer or relevant information for a given query [11]. This is particularly important in scenarios where the modality containing the answer cannot be easily inferred from the question alone.

2.2 Related Work

2.2.1 Sparse vs. Dense Representations

TF-IDF (Term Frequency-Inverse Document Frequency) [1] is a popular algorithm used in sparse retrievers for calculating the similarity between two pieces of text. It combines the concepts of term frequency (TF) and inverse document frequency (IDF) to determine the relevance of a word in a query and a context. TF refers to how many words in the query are found in the context, while IDF is the inverse of the fraction of documents containing a certain term. The TF-IDF score is obtained by multiplying the TF and IDF values. For example, if the word *hippocampus* appears in both the query and the context, it will have a high TF score. Additionally, if it is not found in many other documents (i.e., high IDF), its TF-IDF score will be high. Conversely, common words like *the* will have a low TF-IDF score since they appear in many documents. TF-IDF scores are particularly useful for finding sequences that contain the same uncommon words.

Another widely used method in sparse retrievers is BM25 (Best Match 25) [28], which is a variation of TF-IDF. BM25 incorporates additional adjustments to the scoring mechanism. It dampens the score after returning a large number of matches between the query and the contexts, preventing the dominance of long documents with multiple word matches. Moreover, BM25 considers the length of the documents: it normalizes the score, favoring shorter documents over longer ones when they have the same number of word matches. As a result, BM25 is typically favored over TF-IDF due to its ability to handle large collections of documents more effectively while considering the length of the documents in the scoring process [28].

Moving on to dense retrievers, the research work in [13] is quite important for motivating the present work, especially from the text-modality point of view and the usage of contrastive learning, which we expand on in Section 4.2.2. Here the authors address the need for a more efficient passage retrieval method for open-domain QA. The paper proposes a Dense Passage Retriever (DPR) that uses dense representations alone, learned from a small number of questions and passages by a simple dual-encoder framework, to retrieve relevant passages for a given question. The assumption here is that semantically similar words will be closer in the embedding space. This improves on sparse encoding methods such as BM25 or TF-IDF which do not encode the semantics needed to properly learn [13]. The proposed method outperforms traditional sparse vector space models, such as TF-IDF or BM25, by a large margin in terms of top-20 passage retrieval accuracy [13]. The paper also demonstrates that a higher retrieval precision indeed translates to a higher end-to-end QA accuracy, and achieves new state-of-the-art results on multiple open-domain QA benchmarks, while relying solely on textual information.

However, as mentioned in [13], dense and sparse representations are complementary to each other, and such complementarity lies in their strengths and weaknesses. While sparse representations are memory-efficient and interpretable, they might not capture intricate patterns as well as dense representations. On the other hand, dense representations are powerful and capable of capturing complex relationships, but they come at the cost of higher memory requirements and reduced interpretability. In light of this, in our current work we hope that by leveraging multi-modal (image-text-layout) information, we hope that such complementarity is retained and enhanced.

2.2.2 Vision-Language Representation Learning

Regarding the addition of visual modality to these systems, the work proposed in [12] was pivotal. It was the third of a series of approaches to the challenge of encoding and aligning textual and visual information together [38, 37]. To advance progress in the Document AI community and achieve improved results on document understanding tasks, the authors present LayoutLMv3, a pre-trained multi-modal Transformer for Document AI, which redesigns the model architecture and pre-training objectives of LayoutLM [38]. With this work they introduce the use of unified text and image masking pre-training objectives: masked language modeling, masked image modeling, and word-patch alignment, which the models uses to reconstruct the masked word tokens and image patches simultaneously. The key aspect here is that LayoutLMv3 does not rely on a pre-trained CNN or Faster R-CNN backbone to extract visual features, significantly saving parameters and eliminating the need for region annotations [12]: here each word is mapped with the image patch visually representing the word. The LayouLMv3 model will be used in the present work both for training and evaluation of the different modalities. For additional information and motivation behind this choice, please refer to Chapter 4.

Another pivotal work in the progress of learning new visual concepts directly from raw text only was CLIP [25]. The idea here is to use the semantic information encoded in text in order to inform and support perception learning, leveraging natural language as a training signal used for supervision. They showed that this approach has three main advantages: (1) easier to scale because there is no need for gold-labels, (2) connecting the textual and visual representations enables zero-shot learning and (3) free text usable for supervision is widely available on the web (i.e., image and captions). Published one year later, an extension of CLIP was the work reported in BLIP [19]. Here the authors address the issue of using web datasets for vision-language learning due to the prevalence of noise in web texts. The article proposes a new method called *CapFilt* that utilizes web datasets in a more effective way. Also, the proposed multi-modal mixture of encoder-decoder model offers more flexibility and better performance on a wide range of downstream tasks, while keeping the pre-training simple and efficient [19].

The research in [16] introduces a method for multi-modal retrieval of relevant texts and tables based on questions, given that some questions require information from tables. To address this, the paper presents a method that encodes texts, tables, and questions into a single vector space. To assess the method, authors create a new dataset by combining text and table datasets from prior work. Different encoding schemes, including dense vector embeddings from transformer models and sparse embeddings like TF-IDF and BM25, are compared. The results indicate that dense vector embeddings outperform sparse ones on most evaluation datasets [16]. Their approach employs dense vectors to capture semantic relationships and overcome the limitations of sparse methods like TF-IDF and BM25, which is a motivation for using dense representations in our present project.

Universal Vision-Language Pretraining

To our knowledge, the work in [20] is the only one that moves in our desired direction. The paper introduces Universal Vision-Language Dense Retrieval (UniVL-DR), which aims to establish a unified model for multi-modal retrieval. This approach encodes queries and resources from various modalities into a shared embedding space, facilitating the search for candidates from different sources. To achieve this, UniVL-DR introduces two techniques: a universal embedding optimization strategy that employs modality-balanced hard negatives to enhance the embedding space, and an image verbalization method that bridges the gap between image and text modalities [20].

The authors emphasize that while search engines have traditionally focused on textual data, the growing demand for multimedia content necessitates the incorporation of multi-modal information, which is a common motivation with our work. To address the challenge of merging results from diverse modalities, UniVL-DR seeks to build an end-to-end model that directly maps queries and multi-modal resources into a unified embedding space for retrieval [20].

The paper's experiments compare UniVL-DR with various baseline models, and the experiments demonstrate that UniVL-DR outperforms other models in multi-modal retrieval tasks, achieving substantial improvements in ranking and recall of relevant documents. The modality-balanced hard negative sampling strategy employed by UniVL-DR is highlighted as a key factor in its effectiveness, as it mitigates modality bias during training and enhances modality disambiguation [20]. The image verbalization methods proposed by the authors further enhance the text representations of image documents, in aid of the process of bridging the gap between textual and visual modalities and achieving better retrieval results.

2.2.3 Document VQA

Visual Question Answering (VQA) requires models to understand the semantic content of both the image and the question posed in natural language, and to reason about the relationship between them in order to generate (or extract) an accurate answer. The task of VQA applied to document images is particularly challenging due to the diversity in types of document layouts and content. Tables and graphs, for example, require different types of visual processing than running text (e.g., paragraphs) and titles may need to be recognized as distinct components. VQA is assumed to leverage information such as document structure and metadata (e.g., bounding-box coordinates of words and ROIs within the document image) to improve retrieval performance.

One of the first research efforts with regards to Document VQA was DocVQA [22], which presents a large-scale dataset of almost 13K of document images of varied types and content, over which 50K questions and answers were defined. The dataset was designed for the VQA task on document images. The authors highlight that answering questions in the DocVQA dataset requires reading systems to not only extract and interpret the textual content of the document images but also exploit numerous other visual cues including layout, non-textual elements, and style [22].

However, efforts such as DocVQA did not account for scenarios where the input documents may consist of many pages. Existing datasets and methods for DocVQA focus on single-page documents, which is far from real-life scenario. Therefore, researchers in [34] proposed MP-DocVQA, which is designed for Multi-Page Document Visual Question Answering and aimed at extending single-page DocVQA to the more realistic multi-page setup. The article also proposes a new hierarchical method called Hi-VT5, based on the T5 architecture [26], that overcomes the limitations of current methods to process long multi-page documents [34]. The proposed method is based on a hierarchical transformer architecture where the encoder summarizes the most relevant information of every page, and then the decoder takes this summarized information to generate the final answer [34]. This aligns with the present work on the aim of retrieving possible ROIs out of a larger amount of possible multi-modal contexts.

VisualMRC

Another important research work that stands at the base of this thesis is the VisualMRC. In this article the authors introduce the development of a new task called VisualMRC, which involves reading and comprehending texts given as a document image [33]. The task is decomposed into two sub-tasks: Region-of-Interest (ROI) detection and Optical Character Recognition (OCR). The ROI detection sub-task involves detecting a set of ROIs in an image, where each ROI consists of a bounding box and a semantic class label (e.g., a footer or the caption of a graph, along with their visual coordinates relative to the source document image). The OCR sub-task involves extracting word-level information from the document image along with layout coordinates (bounding-boxes) and confidence scores for each word. The article proposes a model consisting of sub-modules for ROI detection and OCR, and a main module for visual MRC. The main module uses a Transformer architecture and maps an input sequence to a sequence of embeddings, which is passed to the encoder. The input sequence is formed from the tokenization results of the concatenation of a question and OCR words in ROIs. The article also provides a dataset (analyzed in Section 3.1) that includes ground-truth ROIs annotated by humans, and OCR words for each ROI as the outputs of the sub-tasks, as well as relevant ROIs that are required to answer each question. This dataset is very useful to allow models' learning ability based on visually-relevant text-aligning information, such as ROIs and word-level bounding boxes.

The proposed model is an extension of pre-trained encoder-decoder models like BART and T5, which integrates comprehension of visual layout and document content, retaining its pre-trained natural language generation (NLG) abilities. The core main module manages input sequences formed by concatenating questions and OCR words within regions of interest (ROIs), which are then translated into embeddings by the encoder. In this work the authors specify diverse input embeddings,including token, position, segment, location, and appearance embeddings. Token embeddings characterize individual tokens for language representation, while position embeddings encode precise positions. Segment embeddings indicate token classes, offering structural insight. Location embeddings show relative token positions based on bounding box coordinates, and appearance embeddings enhance with visual attributes from ROIs and OCR tokens via a Faster R-CNN model [33]. Another crucial and interesting aspect of this work is saliency detection. A saliency loss mechanism guides token determination, aligning OCR tokens and answers to create pseudo reference labels. The main module's training follows a multi-task approach, concurrently minimizing negative log-likelihood loss and saliency loss through a hyper-parameter. These capabilities (comprehending visual layout, document content, and NLG) empower contextual, multi-modal understanding and human-like language generation.

2.2.4 Uni vs. Multi modality

As far as the comparison of uni-modal and multi-modal approaches to open-domain VQA concern, various research was carried out. For example, in [32] the authors underline the importance of multi-modal encoding mechanisms for answering complex questions that require integrating information across free text, semi-structured tables, and images. The authors demonstrate the necessity of a multi-modal, multi-hop approach to solve their task. Although their multi-hop model, *ImplicitDecomp*, substantially outperforms a strong baseline over cross-modal questions, they show that it still lags significantly behind human performance [32]. Therefore, the authors suggest that multi-modal encoding mechanisms are crucial for improving the performance of open-domain visual question answering models.

WebQA

Another very interesting and useful dataset was considered but eventually unused for this work is WEBQA [4], which focuses on scaling VQA to an open-domain and multi-hop context, mirroring the way humans perform web searches and information retrieval. In this paper the authors highlight the limitations of existing QA systems that often ignore the knowledge present in images and treat the web as a text-only source, thus they emphasize the need for unified multi-modal reasoning models that can answer questions regardless of the source modality. WEBQA dataset includes both image-based and text-based questions, requiring models to perform retrieval, aggregation, reasoning, and natural language generation [4]. The authors also introduce a novel evaluation metric that considers both fluency and accuracy, aiming to capture the challenges of real-world open-domain QA. With this paper they underline the limitations of existing benchmarks, which often focus on template-based or uni-modal approaches, and introduce the need for models that can handle both images and text in an integrated manner. In light of this, they emphasize the need for further research in building unified models that can effectively handle multi-modal information retrieval and reasoning.

While these and other research efforts have shown evidence for the importance of aligning modalities in the context of visual question answering, they also show that such systems do not properly leverage multi-modality yet. Specifically in the context of document VQA, most research directed its focus to text-level understanding, albeit neglecting the layout and structure (ROIs and other components) of the documents [33]. In this work, we will aim to fill this gap by comparing the effectiveness of training on different modalities and their impact on learning properly the alignment of textual and visual information in documents.

Chapter 3

Dataset

As noted by [12], document images are distinct from natural images because they necessitate a precise, detailed alignment between text words and image regions, unlike natural images which do not require it. This alignment relationship is crucial for the successful interpretation and extraction of information from document images, because it allows for the accurate identification and classification of the various components of a document, including text, images, and diagrams. In light of this, only QA datasets retaining certain features are considered for this work. These datasets must be composed of document images which include both visual and textual information, so that comparisons of encoding mechanisms can be fair and relevant. Moreover, the datasets should contain question-answer pairs that are reasonable and pertain to industry-related topics.

To tackle the problem of encoding both visual and textual information in a way that is aligned when passed into a learning model, visually relevant information must be present. For this, the datasets should provide annotations pertaining to the visual components of the document image, such as bounding-box coordinates of ROIs and individual words.

Regions-of-Interest

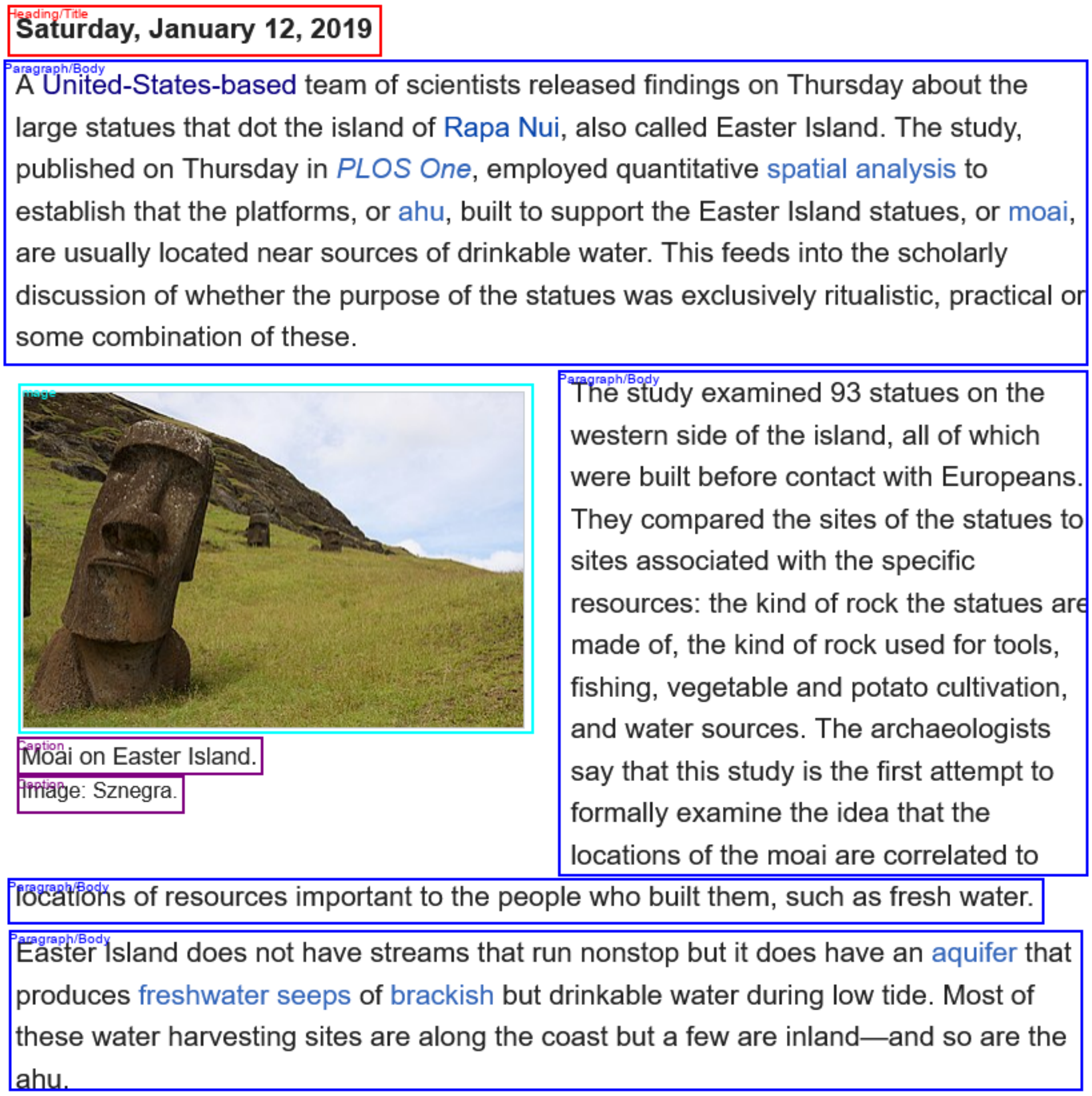
The task of document layout analysis is very similary to other Computer Vision tasks such as image segmentation and object detection, and they all result in dividing an input image into meaningful portions, called Regions-of-Interest (ROIs). In this line of work, such ROIs refer to specific areas or regions within a document image that are semantically meaningful and contain valuable information, i.e., tables, paragraphs or captions. In the context of multi-modal retrieval of regions from a document image, ROIs play a crucial role in identifying and categorizing various components within the document. These components may include headings or titles, and subtitles, bodies of text (i.e., paragraphs), pictures and captions, as well as tables, graphs and the data found in them.

By leveraging ROIs annotations (expressed as bounding box coordinates over the pixel values), systems can efficiently retrieve relevant content from the document image. ROIs facilitate tasks like QA, IR, and content summarization by providing contextually significant segments of text and visual elements. Additionally, we expect it to enhance the performance of our multi-media retrieval system, enabling users to access specific and pertinent information within the document, which is the goal of the present project.

3.1 VisualMRC Dataset: Exploratory Data Analysis

For the purpose of training the multi-modal retriever, we decided to use the VisualMRC dataset [33]. This is a very suitable dataset for the purpose of this project for quite a few reasons, which we outline below. We present an exploratory data analysis and examination of the dataset's characteristics, including its size, composition, distribution, and other relevant statistical insights. Through this analysis, we aim to gain a deeper understanding of the dataset's properties and shed light on its potential implications for the subsequent experiments.

- **Presence of ground-truth ROI annotations**: The dataset provides ground-truth ROIs annotated by humans and OCR words present in each ROI. These are classified into nine classes: *Heading/Title*, *Subtitle/Byline*, *Paragraph/Body*, *Image*, *Caption*, *List*, *Data*, *Sub-data*, and *Other* (please refer to [33] for an in-depth explanation of each ROI class). The presence of such rich types of labelled content enhances the diversity of the dataset. Refer to Figure 3.1 for an example document image with various ROIs and their bounding box coordinates drawn over.



Saturday, January 12, 2019

A United-States-based team of scientists released findings on Thursday about the large statues that dot the island of Rapa Nui, also called Easter Island. The study, published on Thursday in *PLOS One*, employed quantitative spatial analysis to establish that the platforms, or ahu, built to support the Easter Island statues, or moai, are usually located near sources of drinkable water. This feeds into the scholarly discussion of whether the purpose of the statues was exclusively ritualistic, practical or some combination of these.

Moai on Easter Island.

Image: Sznegra.

The study examined 93 statues on the western side of the island, all of which were built before contact with Europeans. They compared the sites of the statues to sites associated with the specific resources: the kind of rock the statues are made of, the kind of rock used for tools, fishing, vegetable and potato cultivation, and water sources. The archaeologists say that this study is the first attempt to formally examine the idea that the locations of the moai are correlated to

locations of resources important to the people who built them, such as fresh water.

Easter Island does not have streams that run nonstop but it does have an aquifer that produces freshwater seeps of brackish but drinkable water during low tide. Most of these water harvesting sites are along the coast but a few are inland—and so are the ahu.

FIGURE 3.1: Sample document image with ROIs bounding boxes and labels drawn over it.

- **Indication of relevant ROIs**: The dataset includes relevant ROIs that are required to answer each question, providing context for the questions. This is useful for the contrastive learning setup that is explained later in Section 4.2.2.

- **Large and diverse collection of document images**: The dataset contains 10,197 images collected from 35 domains [33] (including science, travel, health, news and many others), licensed under creative commons, with content suitable as a document image, containing machine-printed text, pictures, and at least three natural language sentences in each ROI.

- **Three QA pairs per instance**: The dataset consists of three unique questions and their generative answers for each source image, ensuring a comprehensive set of QA pairs. In Table 3.1 three QA pairs are reported along with the ROI relevant to the answer. Moreover, Figure 3.2 shows the distribution of the QA pairs per each split of the data.

Question	Answer	Relevant ROI
What does the picture show?	The picture shows a Moai on Easter Island.	Image
Where are the statues located?	The statues are located on Rapa Nui, also known as Easter Island.	Paragraph/Body
What is the date mentioned at the top?	Saturday, January 12, 2019	Heading/Title

TABLE 3.1: Example of three QA pairs (relative to the document image in Figure 3.1) with the ROI relevant to the answer.

- **Longer and more unique questions and answers**: In Figure 3.3 we report directly the statistics provided in [33], in comparison to the TextVQA [30] and the DocVQA [22] datasets. As we can see, the average question length in VisualMRC is 10.55 tokens, which is larger than in TextVQA (8.12) and DocVQA (9.49), indicating a more diverse and comprehensive set of questions. The dataset also has a higher percentage (96.3%) of unique questions compared to TextVQA (80.7%) and DocVQA (72.3%). The average answer length in VisualMRC is 9.53 tokens, significantly larger than in TextVQA (1.51) and DocVQA (2.43), suggesting more detailed and informative answers. Additionally, it has a significantly higher percentage (91.82%) of questions with unique answers compared to TextVQA (51.74%) and DocVQA (64.29%).

- **More images, tables and graphs**: 44.8% of the document images in VisualMRC contain picture regions and/or data regions such as tables and charts, providing additional visual context. We assume that this abundance of visual information can serve better the purpose of training a multi-modal retriever as it allows it to leverage and learn from this information.

Figure 3.4 reports the distribution of total ROIs per data split, along with the amount that are relevant to answer the questions. As for the QA pairs, the count changes based on the total amount of samples in the data split (i.e., lower in val and test splits, higher in train split).

Figure 3.5 reports the distribution of all the ROI classes per data split. As we can see, there is quite some imbalance across the classes, with amount of ROIs widely

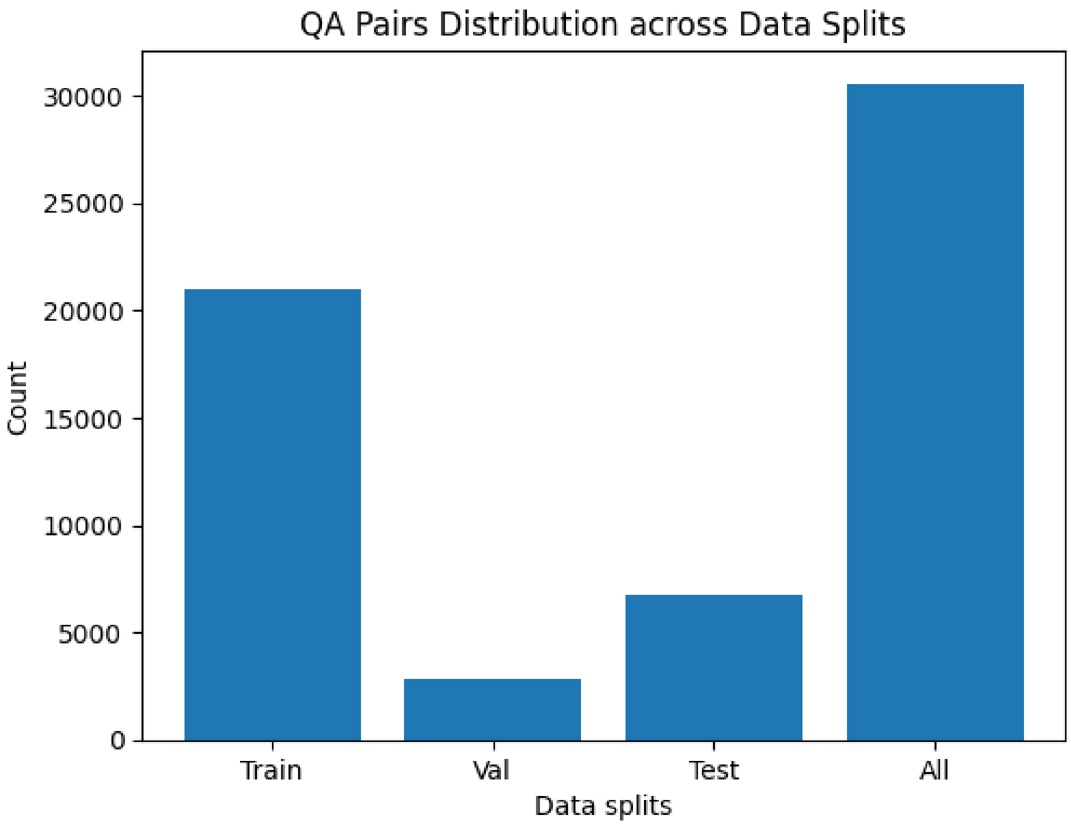
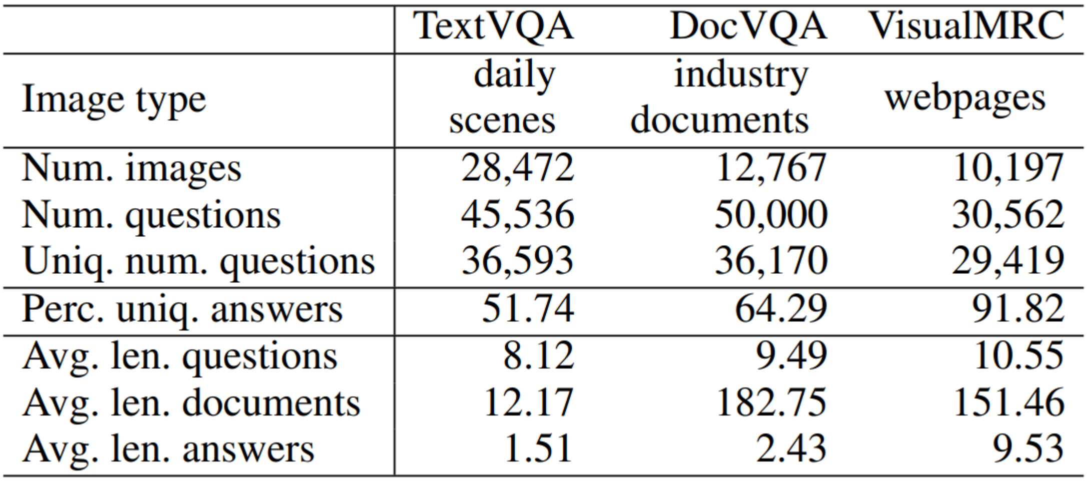


FIGURE 3.2: Histogram distribution of QA pairs across data splits.
2



	TextVQA	DocVQA	VisualMRC
Image type	daily scenes	industry documents	webpages
Num. images	28,472	12,767	10,197
Num. questions	45,536	50,000	30,562
Uniq. num. questions	36,593	36,170	29,419
Perc. uniq. answers	51.74	64.29	91.82
Avg. len. questions	8.12	9.49	10.55
Avg. len. documents	12.17	182.75	151.46
Avg. len. answers	1.51	2.43	9.53

FIGURE 3.3: This table is taken directly from [33] and it compares key statistics of the VisualMRC dataset with the TextVQA [30] and DocVQA [22] datasets. VisualMRC has more unique questions and answers, as well as longer average length of both questions and answers.

varying between classes such as *Paragraph/Body* compared to *Lists* or *Data*. Although this reflects a natural distribution (i.e., documents often contain more text than images, or graphs, or sub-data within the latter), it is sub-optimal to have such imbalance in the training material, when the objective is to properly leverage the not-necessarily

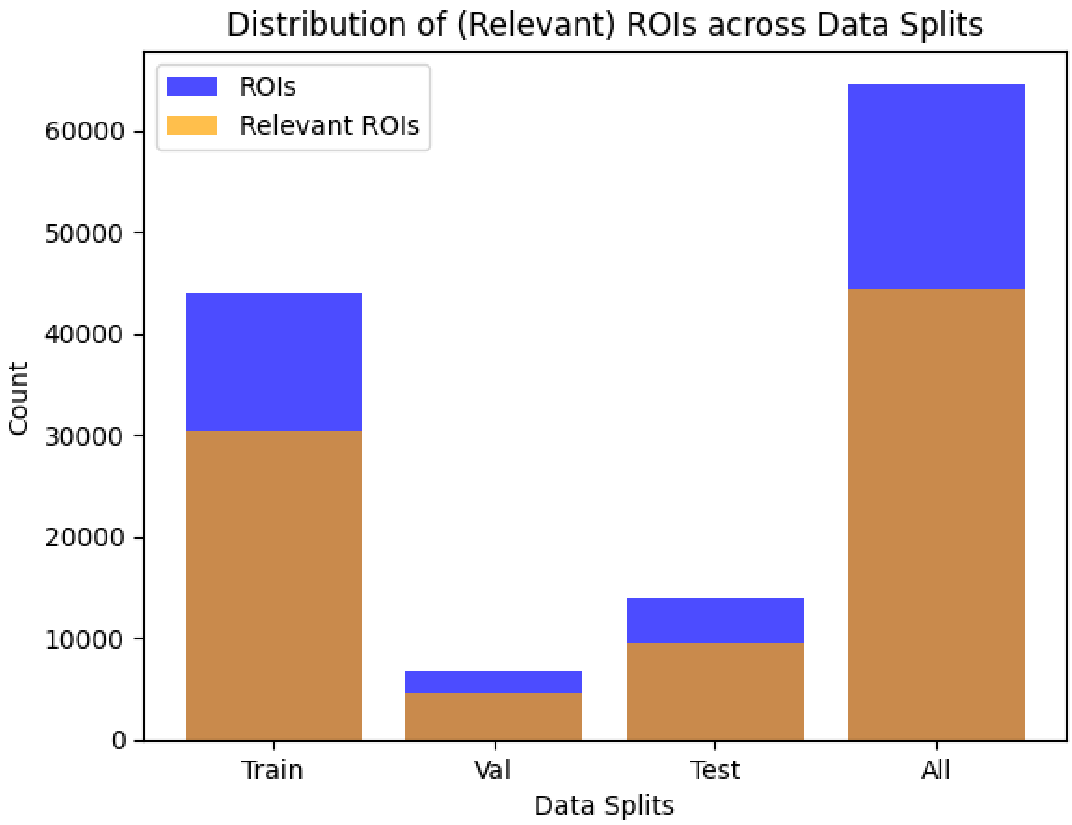


FIGURE 3.4: Histogram distribution of (relevant) ROIs across data splits.

textual information. Nevertheless, the distribution still suggests that enough multimodal information is present and can be leveraged.

In Figure 3.6 we plot the average length of the text for each ROI class and by data split. As expected, classes such as *Paragraph/Body* (mean=54.1 on aggregated data splits) contain on average more words than classes such as *Heading/Title* (mean=5.2 on aggregated data splits) or *Subtitle/Byline* (mean=5.12 on aggregated data splits). Interestingly, we notice that classes such as *Image* or *Sub-Data* contain on average 2.42 and 3.03 words respectively (on aggregated data splits), which show slight inconsistencies in the data collection process. The *Data* class instead correctly reports an average words length of 0 in all data splits.

More interestingly, in Table 3.2 we report example questions for each ROI class. We can see that questions are posed in a way that embed the layout-aware, visually-grounded information to answer the question. This is true especially for ROI classes such as *Image* or *Caption*.

Overall, the VisualMRC Dataset stands out due to its comprehensive annotation of ground-truth ROIs, diverse document image collection, domain diversity, and a wide range of questions and answers, making it more suitable to other datasets like TextVQA [30] and DocVQA [22] in terms of uniqueness and comprehensiveness of content.

3.1.1 VisualMRC pre-processing

In order to ease the process of data extraction for each instance, we decided to create a re-formatted version of the original dataset provided by the researchers. The

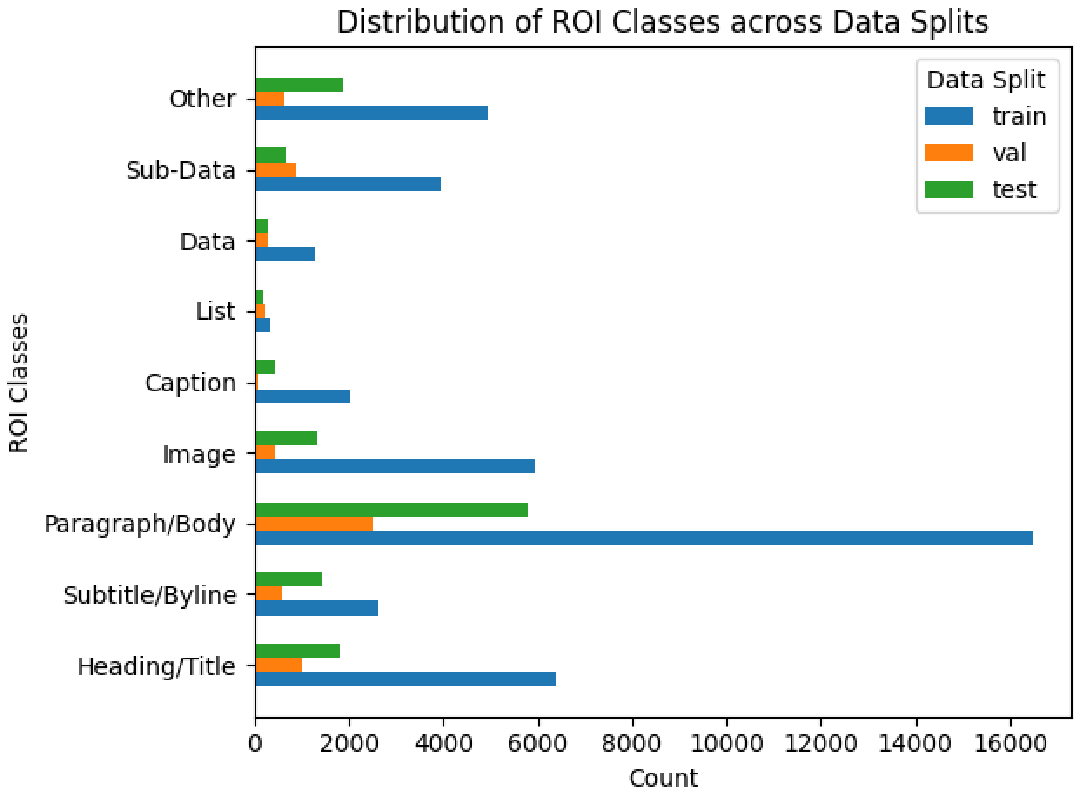


FIGURE 3.5: Bar plot distribution of all the ROI classes per data split.

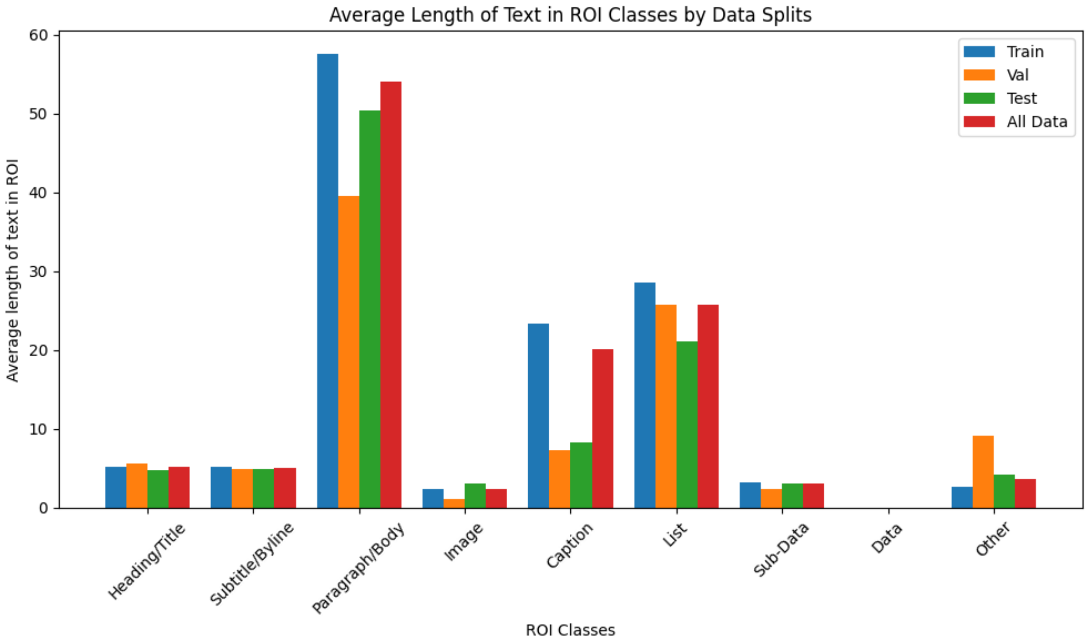


FIGURE 3.6: Average length of texts in each ROI class and by data split.

reasons for doing this are multiple: renaming fields for better readability, deletion of unnecessary fields (e.g., the confidence score of the OCR engine for a specific word), extraction of the words and bounding boxes for each word into lists, removing the need of inefficiently iterating over the nested original format. To do this, we extracted

Heading/Title	**Subtitle/Byline**	**Paragraph/Body**
What is the main issue for real-time programming?	Does the specification have a declaration?	How is the mental ability to travel into the past and future useful to us?
Image	**Caption**	**List**
What building is in the image?	What does the caption say about the picture?	Are the shown authors of the papers all the same?
Sub-Data	**Data**	**Other**
What are the stages of the process of controlling attention during mindfulness practice?	What's the Uruguay Human Development Education Index indicator?	What vitamin is considered key in child malnutrition cases?

TABLE 3.2: Examples of questions for each ROI class.

the data and organized in a tabular format, where each row corresponds to a QA pair and it includes the relevant (positive) and irrelevant (negative) ROIs related to that specific QA pair. It is important to note that we had to account for cases where the OCR information was not present (e.g., for ROIs that do not contain text such as images in the document image), and cases where the amount of total available irrelevant (negative) ROIs was not equal or larger than the amount requested for the specific setting of the number of negative ROIs to use for contrastive learning for each experiment. Specifically for the latter issue, a substantial percentage of samples from the VisualMRC dataset did not contain at least as many negative ROIs as needed for contrastive learning. We initially tried to mitigate this problem by sampling other ROIs from other source documents. However, this naturally resulted in mis-alignment of the bounding box coordinates when applied to the source document. Therefore, in the end, we decided to filter out those instances for which the amount of negative ROIs was less than the required one for the given experiment. This is not an optimal solution, but it locally mitigates the shortcomings of the VisualMRC dataset. These newly formatted .csv files (one for each split) retain the original information (ROIs and QAs) but simplify the access and extraction processes.

Chapter 4

Methodology

In this chapter we explain the dataset used for training, evaluation and testing, along with an exploratory analysis of its statistics, in comparison with other datasets. Moreover, we present our multi-modal retrieval pipeline and explain the tokenization process along with the encoding mechanisms. This allows us to set the ground for the Experiments section which naturally follows this section.

4.1 Processing of Data Inputs

4.1.1 Image resizing

Since the LayoutLMv3 [12] model expects input images to be in a square format of width-height pixel dimensions of 224*224, we first need to resize all the input ROI images to such target dimensions. Figure 4.1 shows ROIs before and after resizing.

4.1.2 Transformation of bounding-boxes of words in ROIs

Now, given that the original bounding box coordinates locating each word on the image are relative to the non-resized image, we also need to transform the bounding-box coordinates to be relative to the resized ROI image. To do this, we use the source and target dimensions and boxes from the input ROI image and transform the coordinates of each word to be relative to the new resized images. Figure 4.2 shows an example of bounding boxes drawn on top of the ROI image after resizing and transformation.

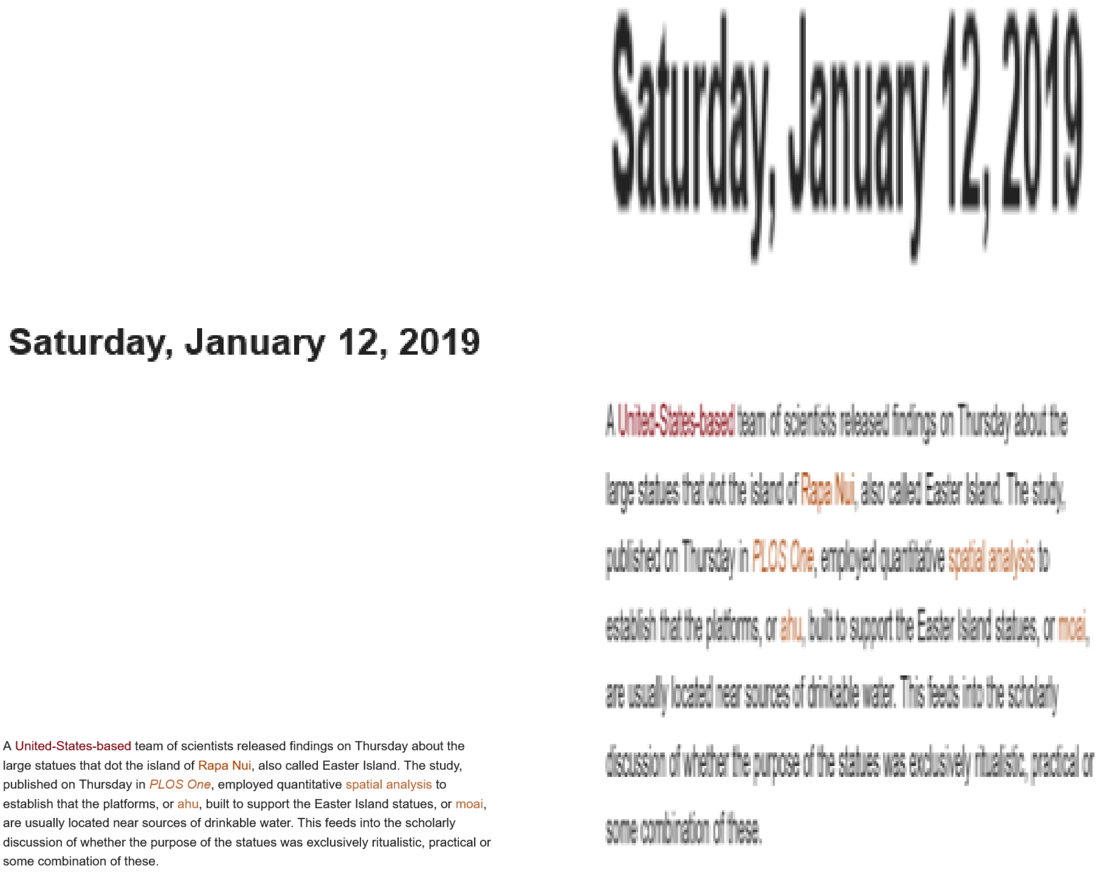
Below in Figure 4.3 the ROI processing diagram combining the steps mentioned above is reported.

4.1.3 Question parsing and tokenization

To parse and tokenize the input questions, we considered using two pre-trained tokenizers implemented on the Huggingface transformers library [35]. On one hand, the BertTokenizer applies end-to-end tokenization on the input sequence, namely punctuation splitting and wordpiece segmentation. On the other hand, the LayoutLMv3Tokenizer is based on the RoBERTatokenizer, which uses Byte Pair Encoding (BPE) and also applies punctuation splitting and wordpiece division. As we mention in the next paragraph, it is also useful for turning the word-level bounding boxes into segment-level bounding boxes, expected by the LMv3 encoder.

4.1.4 Processing of multi-modal ROIs

For processing the multi-modal elements which the ROI contexts (ROI image, ROI text and segment-level coordinates) we use the LayoutLMv3Processor class also from the



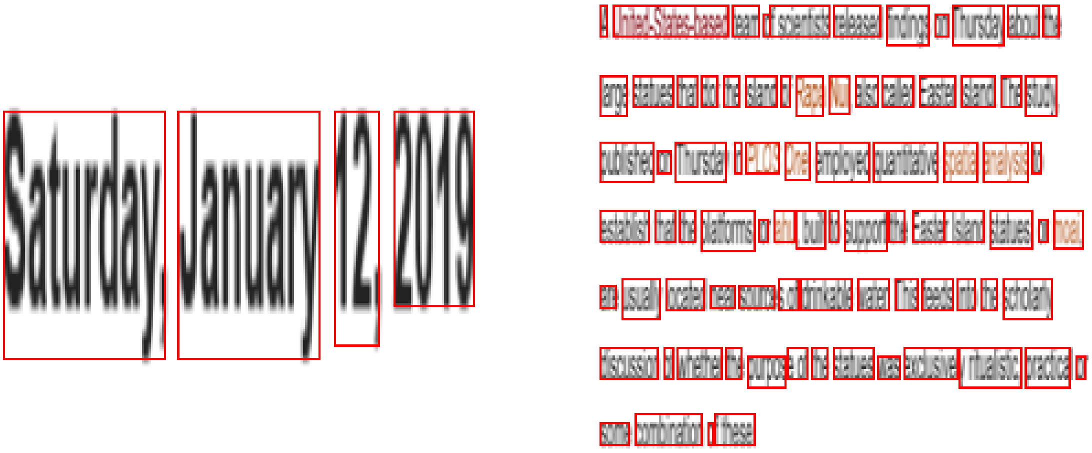
Saturday, January 12, 2019

A United-States-based team of scientists released findings on Thursday about the large statues that dot the island of Rapa Nui, also called Easter Island. The study, published on Thursday in *PLOS One*, employed quantitative spatial analysis to establish that the platforms, or ahu, built to support the Easter Island statues, or moai, are usually located near sources of drinkable water. This feeds into the scholarly discussion of whether the purpose of the statues was exclusively ritualistic, practical or some combination of these.

Saturday, January 12, 2019

A United-States-based team of scientists released findings on Thursday about the large statues that dot the island of Rapa Nui, also called Easter Island. The study, published on Thursday in PLOS One, employed quantitative spatial analysis to establish that the platforms, or ahu, built to support the Easter Island statues, or moai, are usually located near sources of drinkable water. This feeds into the scholarly discussion of whether the purpose of the statues was exclusively ritualistic, practical or some combination of these.

FIGURE 4.1: An example title (top) and paragraph (bottom), before (left) and after (right) image resizing to the target width-height of 224*224 pixels.

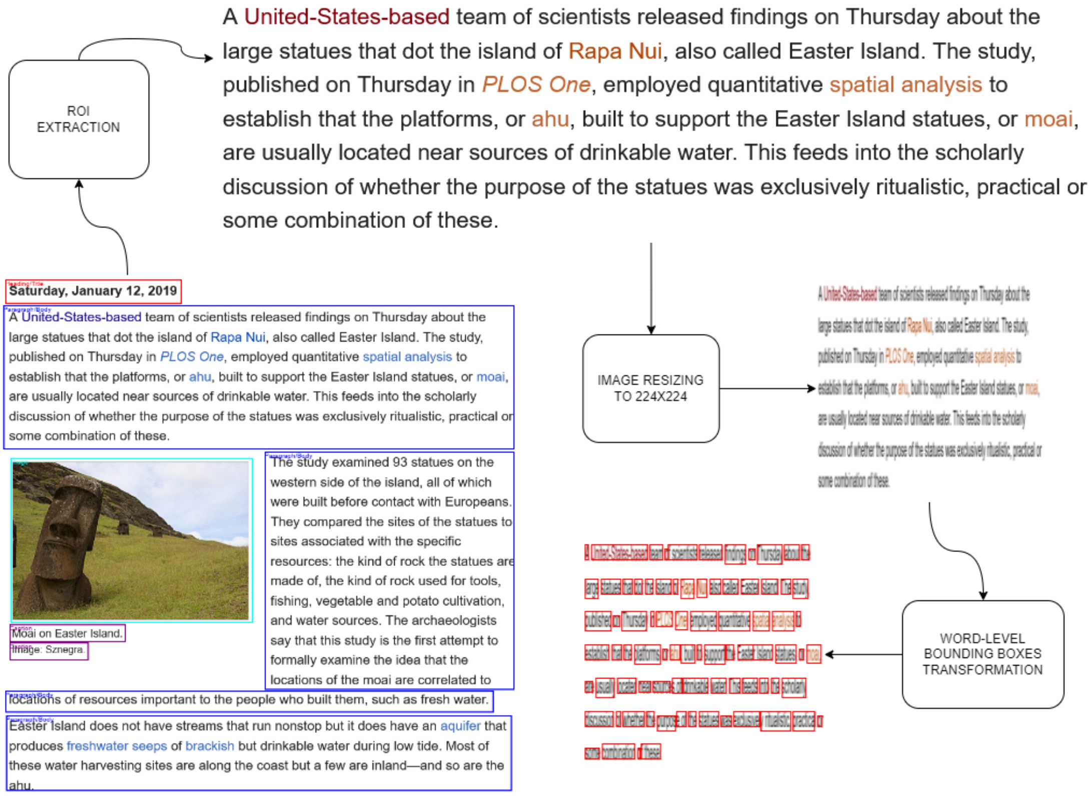


Saturday, January 12, 2019

A United-States-based team of scientists released findings on Thursday about the large statues that dot the island of Rapa Nui, also called Easter Island. The study, published on Thursday in PLOS One, employed quantitative spatial analysis to establish that the platforms, or ahu, built to support the Easter Island statues, or moai, are usually located near sources of drinkable water. This feeds into the scholarly discussion of whether the purpose of the statues was exclusively ritualistic, practical or some combination of these.

FIGURE 4.2: An example title (left) and paragraph (right) after the bounding boxes are transformed to be relative to the new resized ROI images.

Huggingface transformers library [35]. This processor combines a LayoutLMv3 image processor and a LayoutLMv3 tokenizer into a single processor, which is useful because



A United-States-based team of scientists released findings on Thursday about the large statues that dot the island of Rapa Nui, also called Easter Island. The study, published on Thursday in *PLOS One*, employed quantitative spatial analysis to establish that the platforms, or ahu, built to support the Easter Island statues, or moai, are usually located near sources of drinkable water. This feeds into the scholarly discussion of whether the purpose of the statues was exclusively ritualistic, practical or some combination of these.

Saturday, January 12, 2019

A United-States-based team of scientists released findings on Thursday about the large statues that dot the island of Rapa Nui, also called Easter Island. The study, published on Thursday in *PLOS One*, employed quantitative spatial analysis to establish that the platforms, or ahu, built to support the Easter Island statues, or moai, are usually located near sources of drinkable water. This feeds into the scholarly discussion of whether the purpose of the statues was exclusively ritualistic, practical or some combination of these.

Moai on Easter Island.

Image: Sznegra

The study examined 93 statues on the western side of the island, all of which were built before contact with Europeans. They compared the sites of the statues to sites associated with the specific resources: the kind of rock the statues are made of, the kind of rock used for tools, fishing, vegetable and potato cultivation, and water sources. The archaeologists say that this study is the first attempt to formally examine the idea that the locations of the moai are correlated to locations of resources important to the people who built them, such as fresh water.

Easter Island does not have streams that run nonstop but it does have an aquifer that produces freshwater seeps of brackish but drinkable water during low tide. Most of these water harvesting sites are along the coast but a few are inland—and so are the ahu.

FIGURE 4.3: This diagram reports the ROI processing steps that are account for each ROI image from extraction using the provided bounding boxes, to image resizing and the transformation of the word-level bounding boxes after resizing to new target size.

it offers all the functionalities needed to prepare the multi-modal data in a format suitable for the encoder. It first leverages the LayoutLMv3ImageProcessor to extract patch-level image features (resulting in arrays of pixel-values). Here we disable the resizing function (since we already account for it), but we allow rescaling of the input image by a factor of 1/255, which means dividing each pixel value in the image by 255. This process is commonly used to normalize pixel values from the original range of 0 to 255 to a new range of 0 to 1. By performing this rescaling, the pixel values are transformed to a normalized scale, which aids the stable and effective training of neural networks. The specific technique of rescaling pixel values to the range [0, 1] is covered as a fundamental step in preparing data for deep learning models [9].

Moreover, LayoutLMv3 employs linear patches for image embeddings, which serves to mitigate the computational bottleneck of CNNs and eliminate the requirement for region supervision during the training of object detectors [12]. The processor then uses the LayoutLMv3Tokenizer to turn words and layout coordinates into input ids and attention masks for each token and bounding boxes. Also here we disable the OCR option, given that we already have this information available. Important to note: for LayoutLMv3, in the process of tokenizing an OCR word into sub-word tokens, the bounding box coordinates of a sub-word token remain consistent with those of the entire word, following the approach established in the LayoutLM predecessor system [38].

4.2 Modeling Multi-Modal Retrieval

Multi-modal retrieval of regions of interest from document images is a complex task that involves finding relevant information within a document image in response to a given question. To accomplish this, we employ a bi-encoder setup, which means using two separate encoders to convert input data into meaningful representations. In this setup, the input consists of a question and ROIs within the document image, which are described by pixel values, words, and layout coordinates. The latter are collectively referred to as *context*. The goal is to retrieve k amount of contexts which the model scores as relevant or not to answer the question. In Figure 4.4 we report the diagram of the pipeline outlining the steps of the multi-modal retrieval system.

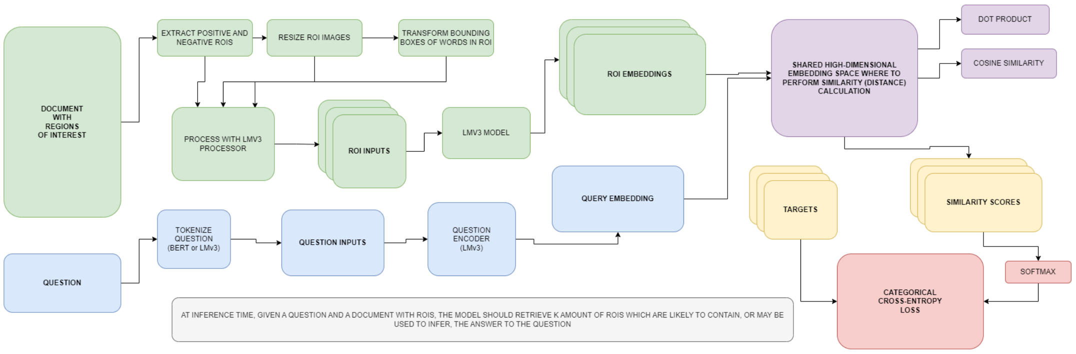

DOCUMENT WITH REGIONS OF INTEREST
EXTRACT POSITIVE AND NEGATIVE ROIS
RESIZE ROI IMAGES
TRANSFORM BOUNDING BOXES OF WORDS IN ROI
PROCESS WITH LMV3 PROCESSOR
ROI INPUTS
LMV3 MODEL
ROI EMBEDDINGS
SHARED HIGH-DIMENSIONAL EMBEDDING SPACE WHERE TO PERFORM SIMILARITY (DISTANCE) CALCULATION
DOT PRODUCT
COSINE SIMILARITY
QUESTION
TOKENIZE QUESTION (BERT or LMv3)
QUESTION INPUTS
QUESTION ENCODER (LMv3)
QUERY EMBEDDING
TARGETS
SIMILARITY SCORES
SOFTMAX
CATEGORICAL CROSS-ENTROPY LOSS
AT INFERENCE TIME, GIVEN A QUESTION AND A DOCUMENT WITH ROIS, THE MODEL SHOULD RETRIEVE K AMOUNT OF ROIS WHICH ARE LIKELY TO CONTAIN, OR MAY BE USED TO INFER, THE ANSWER TO THE QUESTION

FIGURE 4.4: Diagram displaying the steps of the pipeline for the implementation of multi-modal retrieval system. The document image with ROIs is processed (ROIs extraction, resizing and bounding boxes transformation) and processed using LMv3 Processor, which outputs are then passed through the LMv3 Encoder and the embedding representations are obtained. Similarly, the input question is tokenized using the LayoutLMv3 tokenizer and then passed through the LMv3 Encoder to obtain the question embedding representation (we also experimented with a BERT-based question tokenizer and encoder). The similarity score between each context and query embeddings are then calculated in a shared, multi-dimensional dense embedding space and are then passed through a categorical loss function where are compared with the target labels. At inference time, given a question and a document image with ROIs, the model should retrieve k amount of ROIs which are likely to contain, or may be used to infer, the answer to the question.

4.2.1 Encoding Mechanisms

Encoding and representing textual queries provided by users as well as the contextual information needed to answer the queries may be encoded in two ways. The input question is to be parsed, tokenized and encoded via a uni-modal approach, i.e., language-only. As for the document image used as context, it may be processed and encoded with a multi-modal approach, i.e. parsing both the text present in the document and the document image itself, in the form of coordinates (bounding-boxes) representing regions of interest within the document from which the answer can be reasoned upon, and then extracted and/or generated. As outlined in Section 2.2.4, a wide array of approaches exist for representing text and images in a way that is suitable to be processed by the learning models. For this project, we use the fine-tuned encoders described below to represent the input multi-modal data.

Question Encoder: TinyBERT & LMv3

To encode the input question and obtain a dense vector representation of it we experimented with two pre-trained models: the TinyBERT which is streamed down version of BERT [6], and the LayoutLMv3Model [12]. This approach leverages dense vectors to encode tokens with semantic information, which allows to perform similarity calculation between semantically (dis)similar words and thus to compare their vectors in a shared embedding space. As mentioned in 2.2.1, the assumption here is that semantically similar words will be closer in the embedding space. The BERT huggingface implementation for the question encoder is a transformer outputting pooled outputs as question representations, which refers to the last layer hidden-state of the first token of the sequence ([CLS] token) further processed by a Linear layer and a Tanh activation function [6]. This vector representation as the output (d = 768) refers to the fact that only the vector corresponding to the [CLS] token is used as the representation for the entire input sequence. This allows for a fixed-size representation that summarizes the semantic information of the entire input sequence, i.e., it serves as a compact representation that can be used for similarity calculation between questions and passages in the retrieval process. As for the LayoutLMv3Model, more information is provided in the next paragraph.

$$Question_{\text{emb}} = \text{QuestionEncoder}(\text{tokenized\_question}) \tag{4.1}$$

Multi-Modal Encoder: LMv3

To encode the multi-modal context, we make use of the pre-trained LayoutLMv3Model from Huggingface Transformers [35]. This is the bare LayoutLMv3 Model transformer outputting raw hidden-states without any specific head on top. This method uses linear patches for image embedding, which is useful as it not only reduces computational requirements but also eliminates the need for region supervision in training object detectors (focusing on high-level features (e.g., structure of tables perhaps) rather than noisy details). Instead of reconstructing raw pixels or region features, it has learnt to reconstruct discrete image tokens of masked patches. This allows the model to capture the essence of the image without getting distracted by irrelevant information. This model learns to reconstruct masked word tokens from the text modality and symmetrically reconstruct masked patch tokens from the image modality [12]. To obtain the target image tokens, a discrete Variational Auto-Encoder [15] is used to generate latent codes.

$$ROI_{\text{emb}} = \text{ROIEncoder}(\text{processed\_ROI}) \tag{4.2}$$

By directly using raw image patches from document images, LayoutLMv3 jointly learns image, text, and multi-modal representations with unified Masked Language Modeling, Masked Image Modeling, and Word Patch Alignment objectives. The Word Patch Alignment objective is particularly noteworthy as it predicts whether the corresponding image patch of a text word is masked. Important to note: the LMv3 encoder adopts segment-level layout positions, which is an important difference from [38] and [37]. This is because it is based on the assumption that if words appear in the same segment then they are likely to carry very similar meaning and therefore they ought to be represented with common 2D layout positions. Overall, LayoutLMv3 offers a comprehensive solution for multi-modal representation learning by combining innovative techniques for image embedding, pre-training objectives, and cross-modal alignment.

Shared Embedding Space and Similarity Calculation

The output of the encoders ougth to capture the underlying semantics and characteristics of the question and contexts. Thus, we map them in a shared embedding where similarity calculations can be applied. The similarities between the question and the context vectors are then calculated, using two common measures: dot product and cosine similarity. These measures provide a quantitative way to assess how similar the question and context are in the shared cross-modal embedding space. If the vectors are more similar (closer in the embedding space), it suggests that the context may contain relevant information for answering the question.

$$Question_{\text{emb}} \cdot ROI_{\text{emb}} = \sum_{i=1}^{n} Question_{\text{emb}} i \cdot ROI_{\text{emb}} i \tag{4.3}$$

4.2.2 Training Procedure and Contrastive Learning

In this subsection we report the details regarding our approach for training process, along with its objective and the choice of the loss function.

Training objective

The training objective is to optimize the shared cross-modal representation such that the positive contexts are more similar to the question in the embedding space compared to the negative contexts. In other words, the system is expected to learn to differentiate between relevant and irrelevant ROIs for a given question, i.e., which regions of a document image it's relevant to the answer. By iteratively adjusting the model's parameters based on the contrastive learning process, we expect the system to learn to generate meaningful embeddings that facilitate effective retrieval of ROIs.

Contrastive learning

To train the system, a technique known as contrastive learning is used. This involves presenting pairs of positive (relevant) and negative (irrelevant) contexts. The positive context examples are ROIs containing information that is indeed relevant for answering the question. On the other hand, negative context examples are ROIs that do not contain the required information. This technique offers advantages for the task at hand, given that the positive ROI holds essential information and the negative ROIs lack pertinent data. First, we expect it to enable the system to discern subtle distinctions which differentiate meaningful ROIs from their irrelevant counterparts, accomplishing so by attending to visual and semantic cues. Consequently, we expect this capability to empower the system to unravel the complexity of document images, enabling precise localization of ROIs necessary for accurate answers. Second, by iteratively adjusting encoder parameters via the contrastive learning process, the system is expected to generate embeddings that encapsulate rich semantic and visual representations. These embeddings act as concise and meaningful abstractions which should optimize the retrieval of ROIs.

Categorical cross-entropy loss function

The optimization process involves minimizing the categorical cross-entropy loss function, which is defined as the negative log likelihood of the positive ROI. This entails calculating the negative logarithm of the likelihood function, indicating how likely the model believes a given ROI to be positive based on the given question. This comparison involves evaluating the predicted probability distribution against the actual probability distribution (i.e., the target labels). Ultimately, the optimization aims to minimize the loss function by reducing the negative log likelihood associated with the positive ROI. In Equation 4.4, *y_i* refers to the binary label for the positive ROI *i*, while *f_i* represents the score (logit) associated with ROI *i*; the sigma represents the sigmoid function which is used to map the logits to probabilities.

$$\text{Loss} = -\sum_i [y_i \cdot \log(\sigma(f_i)) + (1 - y_i) \cdot \log(1 - \sigma(f_i))] \tag{4.4}$$

4.2.3 Baseline Models

We evaluate the results of our trained model against the following baselines. For the text-only scenario, we consider the LayoutLMv3 model we train on only the question inputs and the texts of the ROIs . For the vision only scenario, we consider the LayoutLMv3 model we train on only the question inputs and the ROIs pixel values. For the multi-modal scenario, we consider the LayoutLMv3 model we train on all the information present in the ROIs (text + image + layout).

4.2.4 Evaluation Metrics

In order to evaluate the performance of our trained models, we use the following evaluation metrics on the test set of the VisualMRC dataset: Normalized Cumulative Discounted Gain (NCDG), Mean Reciprocal Rank @ k, and Recall @ k. These are standard metrics used to evaluate IR systems, and we report them as in [20]. Other metrics we compute but do not report on are precision and hit-rate.

Chapter 5

Experiments

In our setup, we train various models under different configurations, which we explain below. We ran experiments both on a local laptop leveraging the GPU and then moved to a GPU cluster hosted on the cloud to able to increase the amount of data used for training. We experimented with using a separate BERT-based tokenizer and encoder for the input questions, but unfortunately the available resources did not allow us to properly train the models and collect useful results.

5.1 Hyperparameters

For the various experiments, we tweak a few important hyperparameters: the number of negative ROIs for contrastive learning, the effective batch size and, most importantly, the modality on which they are trained. Important to note is that we use accumulation of the gradients, which allows to accumulate the gradients over all the mini-batches after the forward pass, and then normalize them by the amount of steps. This technique allows to increase the effective batch size while keeping the same computational overhead. The assumed relation between the number of negative ROIs and performance is directly proportional: as we add more negative ROIs, the system should have more information to learn the difference in relevance between ROIs. The amount of total samples is used as an experimental value which allowed to run some basic experiments locally. For the modality hyperparameter, we specify the values below. For all the experiments we linearly warmup the learning rate with a warmup ratio of 0.1. Given the memory constraints, we keep an effective batch size of 32 and can only experiment by including up to 3 negative ROIs for each sample. For every experiment, we train for a maximum of 10 epoch, using the early stopping callback with a patience of 3 epochs (i.e., if the score does not improve after 3 epochs, training stops).

5.2 Modalities

5.2.1 Text-only

For the text-only scenario, we tokenize the input question and the text present in the positive and negative ROIs. This results in a batch containing the question input ids and attention masks and ROIs input ids and attention masks, along with the ground truth labels (a tensor of a single 1 and as many 0s as the number of negative ROIs). We then pass this information accordingly to the question encoder and the ROIs encoder, which output the embedding representations that we can use to perform similarity calculation. The embeddings of the ROIs in this case only consider the tokens present in the ROIs, and not the pixel values nor the bounding box coordinates of each token.

5.2.2 Vision-only

For the vision-only scenario, we tokenize the input question and we only extract the features directly from the ROI image. This results in a batch containing the question input ids and attention masks and ROIs pixel values, along with the ground truth labels. We then pass this information accordingly to the question encoder and the ROIs encoder, which output the embedding representations that we can use to perform similarity calculation. The embeddings of the ROIs in this case only consider the pixel values of the ROI images, and not the text contained in them nor the bounding box coordinates of each token.

5.2.3 Multi-Modal

For the multi-modal scenario, we tokenize the input question and we extract the features directly from the ROI images, along with the text present in the ROIs and the bounding box coordinates of each token in that text. This results in a batch containing the question input ids and attention mask, as well as the ROIs pixel values, input ids and attention masks, and segment-level bounding boxes (as in [12]), along with the ground truth labels. We then pass this information accordingly to the question encoder and the ROIs encoder, which output the embedding representations that we can use to perform similarity calculation. The embeddings of the ROIs in this case consider all the vision and language information.

Chapter 6

Results & Discussion

6.1 Results

Here we report the quantitative results of the trained models evaluated on the test set of the VisualMRC dataset. The parameters for the various runs are reported in Chapter 5 and we discuss the results below in Section 6.2.

Important aspects to mention are the fact that we were not able to run the experiments using the bi-encoder setup and also we were not able to iteratively increase the number of negative ROIs per each sample because of memory constraints.

6.1.1 Modalities results

In Table 6.1 below we report the results of the models trained on text-only data, using LayoutLMv3 to tokenize and encode both the question and the ROIs (uni-encoder setup). Under the same configurations we report the results for vision-only and multi-modal scenarios in Table 6.2 and Table 6.3 respectively. For each scenario, under *configuration*, the abbreviations refer to the following: *bs* is the effective batch size, *nnr* is the number of negative ROIs.

Configuration	NCDG@2	MRR@2	Recall@2
bs=32-nnr=1	**0.2614**	**0.3251**	**0.2711**
bs=32-nnr=2	0.1571	0.1999	0.1675
bs=32-nnr=3	0.1678	0.2030	0.1743

TABLE 6.1: Metrics for the models trained on **text-only** data, in the uni-encoder setup.

Configuration	NCDG@2	MRR@2	Recall@2
bs=32-nnr=1	**0.3957**	**0.5565**	**0.4338**
bs=32-nnr=2	0.2253	0.3161	0.2497
bs=32-nnr=3	0.1769	0.2437	0.1954

TABLE 6.2: Metrics for the models trained on **vision-only** data, in the uni-encoder setup.

6.2 Discussion

This thesis proposed to compare the impact of different modalities on retrieval performance of regions of interest from document images, given a user query. As reported

Configuration	NCDG@2	MRR@2	Recall@2
bs=32-nnr=1	**0.3767**	**0.5219**	**0.4262**
bs=32-nnr=2	0.2404	0.3333	0.2654
bs=32-nnr=3	0.1615	0.2225	0.1751

TABLE 6.3: Metrics for the models trained on **multi-modal** data, in the uni-encoder setup.

in Section 6.1 above, we find that both the vision-only and multi-modal scenarios improve retrieval performance of ROIs, over the text-only scenario. Interestingly, the vision-only scenario results in better performance than the multi-modal, which suggests that including the layout information when training does not help the model in differentiating the ROIs relevance given the query. Another consistent effect we can extract from our results is that adding more negative regions of interest for each sample decreases the retrieval performance, suggesting that contrastive learning negatively impacts performance.

6.2.1 Interpretation of Results

In the context of our research questions, the results allow us to validate the first hypothesis: we were able able to treat multi-modal contexts as passages and use them to inform the reasoning for answering a user question. The results, however, do not allow us to fully validate our second hypothesis: we expected multi-modal information (joint representation of text, image and layout) to improve retrieval performance over uni-modal approaches, but this is not fully the case. We can see that multi-modality does increase performance over the text-only approach, but at the same time the vision-only modality results in better performance than the multi-modal approach. We speculate that this is a result of the resource-intensive and data-demanding LayoutLMv3 model that we use for encoding the joint ROI information, which resulted in more computational expenses. Moreover, we believe that the result of the vision-only scenario improving performance is related to our specific dataset and implementation. There are various ways in which we could have devised the inputs to the model, but the one we chose (explained in Chapter 4) seemed the most logical, although perhaps not the most efficient and computationally viable. Finally, adding layout information does not improve performance because of its high impact on the total sequence lengths that are passed to the model during training. Below we discuss a few strengths and limitations of our approach.

6.2.2 Strengths

One of the strengths of our approach is its ability to efficiently retrieve ROIs from a diverse collection of documents in response to user queries. Even in situations where the dataset's OCR quality posed challenges, our filtering mechanism enabled us to retrieve at least a (sub)-optimal number of ROIs for further processing. This efficiency in ROI retrieval is critical for applications that require rapid and accurate access to relevant visual information within a large document corpus, which is in line with the motivations presented in Section 1.2.2.

The dataset adaptability of our approach is another strength worth noting. While the current experiments were based on the VisualMRC dataset, the framework can be extended to other datasets, such as the WebQA [4] (reviewed in 2.2.4), which offer greater diversity and structure. Ideally, we would always prefer to spent as little time as needed on the pre-processing of training data. However, this is often not the case with real-world data, as in the present project. However, the approach suggests adaptability to a broader range of multi-modal retrieval scenarios beyond our initial dataset.

6.2.3 Limitations

The utilization of contrastive learning in shaping the training process does not appear to offer a promising solution. Despite its potential for superior representation learning and improved retrieval capabilities, as suggested by some prior research [20, 13], our findings indicate that it is not a reliable technique for this multi-modal retrieval task. Even when confronted with the limitations of our dataset, contrastive learning fails to demonstrate its viability as a potent tool for enhancing the model's ability to discern relevant ROIs. This discouraging outcome suggests that further exploration and refinement of contrastive learning techniques for multi-modal retrieval may not yield fruitful results.

Contrastive learning heavily relies on the selection of suitable negative samples for training. We did not explore the utilization of hard negatives, a technique employed in prior research and shown to be useful for learning better representations [20, 13]. Incorporating hard negatives into our training strategy could potentially enhance the model's ability to differentiate between positive and negative ROIs. Further exploration of this technique may be worthwhile for future research in contrastive learning.

One of the limitations of our study was the lack of control over the dataset collection process, particularly concerning the Optical Character Recognition (OCR) quality. As discussed in Section 3.1, the quality of the VisualMRC data, which served as the foundation for our contrastive learning model, varied significantly. This variability posed challenges, as certain samples contained insufficient (negative) regions of interest (ROIs) for effective contrastive learning. To address this issue, we implemented a filtering mechanism to exclude samples that did not include at least a given number of negative ROIs. However, this approach may have inadvertently introduced bias into the dataset, which could impact the model's performance.

Another limitation we encountered was the impact of using different tokenizers and encoders on the model's learning process. Given the available computational resources, we were not able to successfully experiment with training bi-encoder setups, where the embeddings of the question and the ROIs are the output of two different encoders. Moreover, local experiments with a lower amount of data revealed that these variations prevented the model from effectively learning the data, resulting in loss spikes and unexpected retrieval performance values. This suggests the need for more systematic investigations into the compatibility and interoperability of tokenization and encoding methods for such multi-modal learning tasks.

An important aspect to mention is that, upon analyzing qualitatively our dataset, we observed that positive and negative ROIs for a given QA pair were not significantly different. We believe this similarity posed a challenge to our system, as it struggled to effectively differentiate between them during training, and indeed our results validate this belief. Future work could explore strategies to increase the dissimilarity between positive and negative ROIs, potentially through data augmentation or more specialized sampling techniques, like in the original DPR paper [13].

Another aspect worth to mention is that, in our project, we chose to resize all ROI images to a fixed size of 224x224 pixels. While this standardization simplifies processing, it may result in the loss of valuable information, especially for ROIs with varying scales. A potential improvement could involve resizing images to a scale factor tailored to each ROI image, ensuring that no information is lost during the preprocessing stage. This may lead to better representation learning, particularly in scenarios with a wide range of ROI sizes.

Chapter 7

Conclusion

7.1 Summary of Findings

This challenging thesis project aimed to compare the impact of different modalities on the retrieval performance of regions of interest (ROIs) in document images given user queries. In Section 6.1, we discovered that both the vision-only and multi-modal scenarios outperformed the text-only scenario in terms of ROI retrieval. Interestingly though, the vision-only approach showed superior performance compared to the multi-modal approach, indicating that including layout information during training did not enhance the model's ability to discern ROI relevance to the query. Another important finding was that increasing the number of negative ROIs for each sample had a detrimental effect on retrieval performance.

In terms of our research questions, we were able to validate the first hypothesis: treating multi-modal ROIs as passages to inform user question reasoning was successful. However, the results did not fully support our second hypothesis, which expected multi-modal information (combining text, image, and layout) to consistently outperform uni-modal methods. We feel safe to suggest that leaving out layout information for this specific task may in the end result in less computational requirements and better retrieval performance.

In conclusion, our multi-modal retrieval approach exhibits some strengths, including efficient ROI retrieval and scalability to diverse datasets. These strengths, although promising, are counterbalanced by the limitations we have encountered in our project, such as dataset variability and encoding techniques that have hindered our progress. While the challenges were many, they also offer valuable insights into areas for improvement and future research directions. As we navigate this balance between strengths and limitations, it is crucial to consider these factors when interpreting our results, ultimately guiding the development of more effective multi-modal retrieval models in the future.

7.2 Implications and Future Work

A viable and promising way forward for enhancing our multi-modal retrieval approach would be to include the WebQA dataset, as done in [20]. The WebQA dataset offers a wealth of diverse and real-world data, providing a unique opportunity to evaluate our approach in a more challenging and ecologically valid setting. More broadly, we suggest future work in this area to compose a custom-made dataset which preserves the useful characteristis of question and answer diversity of the VisualMRC dataset and the approachability and ease of use of the WebQA dataset. We expect such effort to help gain insights into how model performs under different data conditions and further refine its capabilities.

Recently, at a conference, I found out about an interesting work by Berrios et.al [3], which introduces the LENS framework. Here the idea is to have Large Language Models (LLMs) to solve computer vision tasks, which could open up new possibilities for advancing our multi-modal retrieval approach. LENS introduces innovative techniques by freezing the components of the pre-trained models and using a mix of image features, extracted with CLIP [25] and BLIP [19], which are then used to prompt a Large Language Model (LLM) and generate the answer. This approach has the advantage that no training is required. Therefore, we think that incorporating elements of LENS into our framework may enable us to address some of the limitations we have encountered and enhance the overall effectiveness of our retrieval system.

To gain a more nuanced understanding of our results, a fruitful direction for future research involves categorizing queries based on question types. By dissecting the retrieval outcomes along selected dimensions, such as the types of images or whether they pertain to spatial or temporal dimensions (as discussed in Section3.1), we can uncover patterns and performance variations that might be obscured in an aggregated analysis. This approach can provide valuable insights into the strengths and weaknesses of our model across different query categories, aiding in the development of more specialized retrieval strategies.

An intriguing proposition for enhancing our multi-modal retrieval pipeline is the integration of a keyword matching system early in the pipeline. This system could serve as a preliminary filter, narrowing down the search space based on keyword matches before applying the full multi-modal retrieval pipeline. This approach may also enhance the reliance on specific keywords and help us identify potential spurious correlations. We believe that by uncovering the question types the model relies on when predicting the relevance of ROIs, we can refine our retrieval strategy and improve the model's overall performance.

For a more comprehensive and advanced retrieval methodology, future research could explore the integration of a retrieval augmented generation approach, which is a very recent and promising research field. By combining retrieval techniques with LLMs and natural language generation, we can expand the pipeline's capabilities.

The present research was significantly relevant for the field of multi-modal retrieval, and it hopes to provide valuable insights for enhancing neural network models dedicated to document visual information retrieval. By emphasizing the importance of both visual layout and textual context, it contributes to the expansion of current research efforts, as it seeks to deepen our understanding of the actual usefulness of unified multi-modal approaches, and inspire further progress in this or other directions.

Bibliography

[1] Rajaraman Anand and Ullman Jeffrey David. *Mining of massive datasets*. Cambridge university press, 2011.

[2] Payal Bajaj et al. *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset*. 2018. arXiv: 1611.09268 [cs.CL].

[3] William Berrios et al. "Towards language models that can see: Computer vision through the lens of natural language". In: *arXiv preprint arXiv:2306.16410* (2023).

[4] Yingshan Chang et al. "Webqa: Multihop and multimodal qa". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16495–16504.

[5] Fabio Crestani et al. ""Is this document relevant?... probably" a survey of probabilistic models in information retrieval". In: *ACM Computing Surveys (CSUR)* 30.4 (1998), pp. 528–552.

[6] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[7] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[8] Yifan Du et al. "A survey of vision-language pre-trained models". In: *arXiv preprint arXiv:2202.10936* (2022).

[9] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press, 2016.

[10] Bert F. Green et al. "Baseball: An Automatic Question-Answerer". In: *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*. IRE-AIEE-ACM '61 (Western). Los Angeles, California: Association for Computing Machinery, 1961, 219–224. ISBN: 9781450378727. DOI: 10.1145/1460690.1460714. URL: https://doi.org/10.1145/1460690.1460714.

[11] Darryl Hannan, Akshay Jain, and Mohit Bansal. "Manymodalqa: Modality disambiguation and qa over diverse inputs". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7879–7886.

[12] Yupan Huang et al. "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking". In: *arXiv preprint arXiv:2204.08387* (2022).

[13] Vladimir Karpukhin et al. "Dense passage retrieval for open-domain question answering". In: *arXiv preprint arXiv:2004.04906* (2020).

[14] Wonjae Kim, Bokyung Son, and Ildoo Kim. *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision*. 2021. arXiv: 2102.03334 [stat.ML].

[15] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[16] Bogdan Kostić, Julian Risch, and Timo Möller. "Multi-modal retrieval of tables and texts using tri-encoder models". In: *arXiv preprint arXiv:2108.04049* (2021).

[17] Julian Kupiec. “MURAX: A robust linguistic approach for question answering using an on-line encyclopedia”. In: *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 1993, pp. 181–190.

[18] Cody C. T. Kwok, Oren Etzioni, and Daniel S. Weld. “Scaling Question Answering to the Web”. In: *Proceedings of the 10th International Conference on World Wide Web*. WWW ’01. Hong Kong, Hong Kong: Association for Computing Machinery, 2001, 150–161. ISBN: 1581133480. DOI: 10.1145/371920.371973. URL: https://doi.org/10.1145/371920.371973.

[19] Junnan Li et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. 2022. arXiv: 2201.12086 [cs.CV].

[20] Zhenghao Liu et al. “Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval”. In: *The Eleventh International Conference on Learning Representations*. 2023.

[21] Jiasen Lu et al. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. 2019. arXiv: 1908.02265 [cs.CV].

[22] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. *DocVQA: A Dataset for VQA on Document Images*. 2021. arXiv: 2007.00398 [cs.CV].

[23] Arsha Nagrani et al. “Attention bottlenecks for multimodal fusion”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14200–14213.

[24] John Prager et al. “Open-domain question–answering”. In: *Foundations and Trends® in Information Retrieval* 1.2 (2007), pp. 91–231.

[25] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[26] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2020. arXiv: 1910.10683 [cs.LG].

[27] Pranav Rajpurkar et al. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. 2016. arXiv: 1606.05250 [cs.CL].

[28] Stephen E Robertson and Steve Walker. “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval”. In: *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer. 1994, pp. 232–241.

[29] G. Salton, A. Wong, and C. S. Yang. “A Vector Space Model for Automatic Indexing”. In: *Commun. ACM* 18.11 (1975), 613–620. ISSN: 0001-0782. DOI: 10.1145/361219.361220. URL: https://doi.org/10.1145/361219.361220.

[30] Amanpreet Singh et al. “Towards VQA Models That Can Read”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[31] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. “A literature review on question answering techniques, paradigms and systems”. In: *Journal of King Saud University-Computer and Information Sciences* 32.6 (2020), pp. 635–646.

[32] Alon Talmor et al. *MultiModalQA: Complex Question Answering over Text, Tables and Images*. 2021. arXiv: 2104.06039 [cs.CL].

[33] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. *VisualMRC: Machine Reading Comprehension on Document Images*. 2021. arXiv: `2101.11272 [cs.CL]`.

[34] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. *Hierarchical multimodal transformers for Multi-Page DocVQA*. 2022. arXiv: `2212.05935 [cs.CV]`.

[35] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020. arXiv: `1910.03771 [cs.CL]`.

[36] William A Woods. *Conceptual indexing: A better way to organize knowledge*. Sun Microsystems, Inc., 1997.

[37] Yang Xu et al. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. 2022. arXiv: `2012.14740 [cs.CL]`.

[38] Yiheng Xu et al. "LayoutLM: Pre-training of Text and Layout for Document Image Understanding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. ACM, 2020. DOI: `10.1145/3394486.3403172`. URL: `https://doi.org/10.1145%2F3394486.3403172`.