Master's Thesis

Efficient Textual and Visual Question-Answering Systems for Scientific Documents

LE, Khang Nguyen

Supervisor NGUYEN, Minh Le

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Master Degree)

March, 2023

Abstract

Visual-Question-Answering (VQA) requires a VQA system to answer questions corresponding to visual information. VQA in mathematics abstract diagrams containing abstract objects instead of natural images requires diverse cognitive reasoning skills, posing many challenges to current VQA methods. Language-vision models whose vision features come from image patch tokens or object proposals may not efficiently capture and present the information about the type and number of objects, which is essential for answering the mathematical question. Object detection techniques are important for obtaining object proposals. However, existing object detection models are trained on natural images, and no dataset is available for fine-tuning object detection on abstract objects. This study proposes methods for detecting abstract objects and generating valuable descriptions that can be used to augment the training and inference process of abstract diagram VQA models. The experiments show that existing VQA models benefit greatly from the augmented descriptions. Moreover, Transformer models trained using only the descriptions without any visual information achieve state-of-the-art results in IconQA sub-tasks. Furthermore, the proposed abstract object detection method enables future research in abstract diagram VQA models that use features from object proposals.

Acknowledgement

I would like to express my profound gratitude to my supervisor Professor NGUYEN Minh Le, my second supervisor Professor Satoshi Tojo, and my supervisor for minor research Professor Shinobu Hasegawa for their invaluable guidance, support, and mentorship throughout this journey. Their expert knowledge and critical insights have been essential in shaping the direction of my research and helping me to reach my full potential. Their patience, encouragement, and willingness to offer their time and resources were truly invaluable.

I would like to extend my thanks to the faculty and staff of Japan Advanced Institute of Science and Technology (JAIST), for providing me with the resources, opportunities, and infrastructure necessary to complete my thesis. Their support and guidance have been critical to my academic success, and I am grateful for their dedication to helping me succeed.

I would also like to express my sincere appreciation to the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for their support in the form of the MEXT scholarship. This scholarship allowed me to pursue my academic interests and conduct my research in Japan, providing me with the necessary resources and opportunities to excel. This scholarship will remain a cherished memory and an integral part of my academic journey.

I am deeply thankful to my friends and family for their unconditional love and support throughout this process. Their encouragement and understanding have been a source of strength, and I could not have completed this project without their support.

Finally, I would like to acknowledge all those who have directly or indirectly contributed to this work, including participants in my study and colleagues in the field. Your input and assistance have been invaluable and are greatly appreciated.

Contents

1 Introduction **1**
- 1.1 Background . . . 1
  - 1.1.1 Visual-Question-Answering on Abstract Diagrams . . . 1
  - 1.1.2 Finding information in Scientific Papers . . . 4
- 1.2 Objectives . . . 6

2 Related Works **8**
- 2.1 Question-Answering . . . 8
- 2.2 Visual-Question-Answering . . . 9
- 2.3 Finding information in Scientific Papers . . . 11

3 Proposed Method **13**
- 3.1 Description Augmentation for VQA in Mathematics Abstract Diagram . . . 13
  - 3.1.1 Problem formulation . . . 13
  - 3.1.2 Object Detection for Abstract Diagram . . . 13
  - 3.1.3 Description Generation . . . 16
  - 3.1.4 Description Augmentation for VQA Models . . . 16
  - 3.1.5 Description-based Models . . . 17
- 3.2 Finding information in scientific papers . . . 19

4 Experimentation and Evaluation **22**
- 4.1 Description Augmentation for VQA in Mathematics Abstract Diagram . . . 22
  - 4.1.1 Experimental Setup . . . 22
  - 4.1.2 Results . . . 24
- 4.2 Finding information in Scientific Papers . . . 26
  - 4.2.1 Experimental setup . . . 26
  - 4.2.2 Experimental Results . . . 27

5 Conclusions **35**
5.1 Conclusions 35
5.2 Publications 36
5.2.1 Publications related to the thesis 36
5.2.2 Other publications 36

List of Figures

1.1 Examples in IconQA dataset 3
1.2 Examples in IconQA dataset 4

3.1 Examples of three sub-tasks in IconQA 14
3.2 Description Generation Process 17
3.3 Models for mathematics abstract diagram VQA using only the generated descriptions. No visual information is used. 18
3.4 Retriever-Reader approach on long scientific articles 20

4.1 Accuracy on 13 skill categories in IconQA test set 31
4.2 Examples of description generation on IconQA dataset 32
4.3 Effect of various number of retrieved passages and window size on the accuracy of Retrieval. Contrasting the performance of the Sparse Retriever utilizing TF-IDF and the Cross-encoder (CE) with diverse window size $W \in \{100, 150, 200\}$ 33
4.4 Investigation of the impact of utilizing a sliding window technique on the performance of end-to-end question answering with various reader models 33
4.5 Examination of the effects of various window sizes on the performance of different retriever methods 34
4.6 Evaluation of retrieval techniques at varying levels of retrieved passages. The solid lines indicate accuracy when at least one of the retrieved passages contains the correct answer (One). The dashed lines indicate the accuracy when all retrieved passages contain the correct answer (All) 34

List of Tables

4.1 Accuracy on IconQA test set. Patch-TRM (Reproduced) is from reproducing experiments using the provided code, used for comparison. Patch-TRM (From paper) is taken from the cited paper, used for reference. The proposed description augmentation method is evaluated in two settings: (1) Description augmentation for existing VQA models. (2) Description-based models using only the descriptions without visual information. 22

4.2 Accuracy on IconQA test set of different description-based models 23

4.3 Reasoning skill categories in IconQA 24

4.4 Overall of IconQA dataset 24

4.5 Assessment of the proposed method against current state-of-the-art techniques on the QASPER development set, specifically the extractive question subset 28

4.6 An examination of individual components through removal and substitution 29

Chapter 1

Introduction

1.1 Background

1.1.1 Visual-Question-Answering on Abstract Diagrams

Visual-Question-Answering (VQA) is an important task requiring systems to answer natural language questions corresponding to visual information. Advances in VQA have been witnessed in recent years, with several datasets proposed. The visual information in these datasets can be categorized into natural images and abstract diagrams. While many VQA datasets focus on natural images [3, 83, 18, 26, 24, 73], some work attempts to address the VQA task in abstract diagrams. As diagrams account for a considerable share of the visual world, the types of abstract diagrams in VQA datasets are diverse, including charts and illustrations [28], geometry [47], scientific diagrams [32], and others [39, 50, 33]. Different types of abstract diagram VQA datasets aim to evaluate different aspects of VQA systems. Specifically, abstract diagram VQA in math problems is shown to involve diverse reasoning skills [30] and has a high potential for the development of educational applications.

In the task of Visual Question Answering (VQA), a model is given an image along with a natural language question pertaining to the image and is required to output a response. The objective of VQA is to evaluate a model's capability to comprehend both visual and textual inputs and to utilize that comprehension to produce a coherent and appropriate answer. VQA models typically incorporate a combination of computer vision techniques to comprehend image content and natural language processing techniques, to comprehend the question and generate the answer. The VQA task is considered to be a challenging task due to both visual and textual modalities and their interactions.

VQA has many publicly available datasets for training and evaluating

models, such as VQA v2, COCO-QA, and Visual7W. These datasets contain a large number of images and questions along with their answers. The VQA task has many applications in areas like visual surveillance, autonomous navigation, and human-computer interaction. The VQA task can be used to improve the performance of other tasks, such as image captioning and text-to-image retrieval.

Recent advancements in deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have led to significant improvements in the performance of VQA models. However, there still exists a gap between human-level performance and the performance of current VQA models. Future research in VQA can focus on developing models that can better understand the context, generate more human-like answers, and can generalize to unseen images and questions [3].

Despite being practical in real-world problems, math abstract diagram VQA tasks requiring reasoning and math skills are understudied. Lu et al.[51] proposed IconQA as a dataset for abstract diagram visual reasoning and question answering, with questions inspired by math problems for children [53]. Figure 1.1 and 1.2 shows some examples in IconQA dataset. IconQA has posed many challenges for current VQA methods due to its uniqueness in questions and visual information. Firstly, the questions require various human cognitive skills on different levels, ranging from pre-kindergarten to third grade. Secondly, the visual information contains abstract icons, which are challenging for many vision-language models pre-trained on natural images [46, 9, 14, 37].

VQA task in understanding and reasoning about abstract diagrams is a cognitive task that requires a model to extract and interpret the meaning of abstract diagrams, and use that understanding to answer questions or complete other tasks. This task can be divided into two main components: understanding the meaning of the diagram and reasoning about the diagram. Understanding the meaning of the diagram involves recognizing the various elements of the diagram and their relationships to each other, such as recognizing the different shapes, lines, and colors in the diagram. Reasoning about the diagram involves using the information in the diagram to answer questions or make inferences. For example, in an organizational chart, use the diagram to answer questions about reporting lines or infer the position of an employee in the company hierarchy.

Research in this area has been focusing on developing models that can perform this task using computer vision and natural language processing techniques. These models typically use a combination of CNNs and RNNs to extract visual features from the diagrams and understand the question and then use this information to generate a coherent and relevant answer.

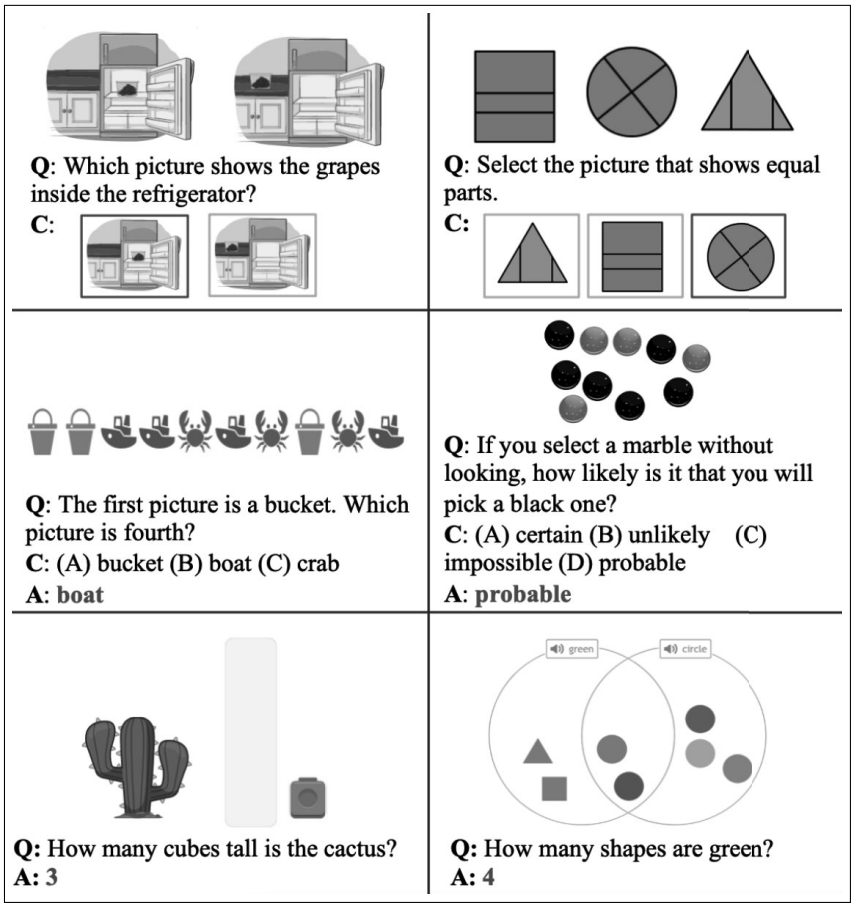


Figure 1.1: Examples in IconQA dataset

However, the task of VQA in understanding and reasoning about abstract diagrams is still considered to be a challenging task, and there is still a gap between the human-level performance and the performance of current models.

Attention-based multi-modal approaches [2, 36, 16, 79] and VQA models based on Transformers [46, 9, 14, 37] have shown promising results on VQA tasks with natural images and not involving cognitive reasoning. However, the performance of these methods is not as efficient when adapting to the VQA tasks of IconQA. To answer questions requiring reasoning and mathematics skills, the model should be aware of the separate entities in the visual context, the number, and the properties of these entities. Language-vision models whose vision features come from image patch tokens [14, 37, 51] may not be suitable for mathematics problems because the objects in the visual context are separated, and the patches might not capture these ob-

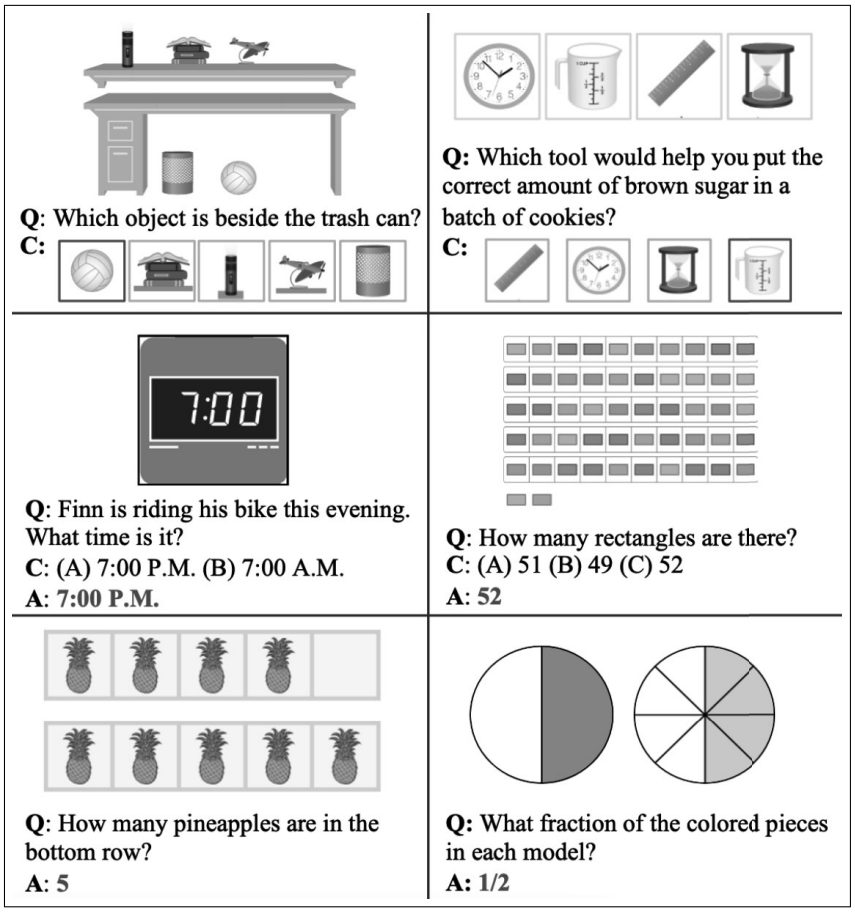


Figure 1.2: Examples in IconQA dataset

jects effectively. VQA models using features from image object proposals [36, 16, 79, 46, 9] relies on effective object detection models like Faster-RCNN [60] to extract object proposals. However, no data is available for fine-tuning object detection models to detect abstract icon objects.

1.1.2 Finding information in Scientific Papers

The ability to extract answers from long scientific documents is a challenging task for intelligent systems. This research examines the application of retriever-reader methods, commonly used in open-domain QA, to this specific task. The proposed approach utilizes a single scientific article as a corpus for retrieval and employs an efficient sliding window technique to improve the pipeline by dividing the article into fixed-size text blocks. The results of experimentation on the QASPER dataset, designed for QA in NLP pa-

pers, show that the proposed method surpasses existing state-of-the-art models, achieving a new state-of-the-art in the extractive questions subset with 30.43% F1.

QA involves utilizing a given question and context to find the corresponding answer within the context. A variety of datasets have been proposed for the QA task, including SQUAD [58], HotpotQA [78], WebQuestions [68], NaturalQuestions [40], and TriviaQA [27]. Dasigi et al. [11] have introduced QASPER, a dataset specifically designed for QA on NLP papers. The dataset requires finding an answer to a question about an academic article, given the complexity and length of the contexts. QASPER comprises four types of answers: extractive, abstractive (free-form answer), yes/no, and unanswerable. This study focuses specifically on the questions that have extractive answers.

QASPER is a unique QA dataset in that it features significantly larger contexts for the questions than other datasets. Specifically, scientific papers are often much longer than the token limit of 512 or 1024 that are commonly used by BERT-like models such as BERT [12], RoBERTa [44], ELECTRA [10], and ALBERT [41]. This presents a challenge for applying these models to QASPER. Dasigi et al. [11] suggested QASPER-LED as a solution for processing long sequences, using Longformer-Encoder-Decoder [4] that can handle 16K tokens without truncation, but this may not efficiently capture the meaning of all tokens.

In recent years, Open-domain QA has attracted a lot of attention in the research community. In this task, the system is only given a question and must find the answer from a vast collection of documents. A common strategy for this task is to use a two-stage approach, which includes a retriever and a reader. The retriever's job is to find relevant documents from the corpus, while the reader's task is to locate the answer within those documents. Numerous methods have been proposed for the retriever stage. These methods can be broadly divided into two categories: Sparse Retrievers and Dense Retrievers. Sparse Retrievers use traditional methods such as TF-IDF and BM25. One of the first approaches to combine these retrieval techniques with neural models for MRC was DrQA [8]. In contrast, Dense Retrievers use dual-encoders to encode the question and document. Notable examples include Sentence-BERT [59], DPR [31], and ORQA [42].

In some respects, the task of finding information in a long scientific paper in QASPER is similar to that of open-domain QA. The use of retrieval techniques in the initial stage of open-domain QA may be beneficial for locating pertinent information within QASPER. To the best of our knowledge, no prior studies have examined the effects of open-domain QA retrieval methods on the QASPER dataset. This paper presents a new approach that adapts the retriever-reader methodologies of open-domain QA to long scien-

tific papers within QASPER by treating each document as a large corpus for retrieval. By doing so, we are able to apply open-domain QA techniques and develop a retriever-reader pipeline that achieves state-of-the-art performance on QASPER.

1.2 Objectives

This research aims to propose efficient VQA systems to address the mentioned challenges in mathematics abstract diagram VQA in textbooks and propose a method for effectively finding information in these documents. To address these challenges in mathematics abstract diagram VQA, we propose a method to generate descriptions for the mathematics problems from the visual information. The descriptions are used to augment the training and inference process of VQA models and guide the models to perform mathematics reasoning. All the code and models of the proposed method are available. The research has the following contributions.

1. We propose a method for generating useful descriptions to augment the training and inference process of abstract diagram VQA models. The descriptions provide essential information for the models to answer mathematics questions. The description augmentation process is model-agnostic and can be applied to any VQA model to enhance its performance. The experimental results show that the performance of existing VQA models can benefit greatly from the description augmentation.

2. We develop a set of abstract diagram question-answering models for mathematics problems using only the generated descriptions without any visual information. The experiments on IconQA show that the models achieve competitive results in *multi-image-choice* sub-tasks and state-of-the-art results in *multi-text-choice* sub-tasks, using only the signals from the descriptions.

3. We proposed an object detection method specially designed for abstract diagrams using Connected-Component Labeling. This method overcomes the lack of abstract-object detection datasets and enables future research in abstract diagram VQA, especially VQA models that use object proposals as the visual features.

4. We proposed a retriever-reader method for finding information in scientific documents and an effective sliding window technique that im-

proves the pipeline. The experiments of this method are conducted on QASPER dataset.

Chapter 2

Related Works

2.1 Question-Answering

Question answering (QA) is a task in natural language processing (NLP) and information retrieval (IR) that aims to provide accurate and concise answers to questions posed in natural language. The goal of QA is to extract relevant information from a wide range of sources, including but not limited to, news, literature, and databases, and present it in a coherent and understandable format. There are two main approaches to QA systems: rule-based and machine learning-based. Rule-based systems rely on pre-defined heuristics and hand-crafted rules to extract information and provide answers. In contrast, machine learning-based systems utilize statistical models to extract information and provide answers.

Recent advancements in deep learning techniques, particularly transformer-based models [12][5][44], have a great impact on the performance of QA systems. These models are trained on large amounts of text data and have the ability to understand the context of the question and generate accurate and coherent answers.

QA systems have many applications, including customer service, personal assistants, and educational systems. The task of QA is considered challenging due to the need for logical reasoning to generate accurate answers. Future research in QA can focus on developing models that can better understand the context, generate more human-like answers, and generalize to unseen questions. In customer service, QA systems can be used to provide quick and accurate answers to customer inquiries, reducing the need for human customer service representatives. By automating the answering process, QA systems can handle a large volume of customer queries and provide consistent and accurate responses. This can lead to improved customer satisfaction

and reduced costs for the company. For personal assistants, QA systems can be integrated into personal assistants such as Siri, Alexa, and Google Assistant to provide users with accurate and personalized answers to their queries. These systems can answer a wide range of questions, from providing weather forecasts to recommending restaurants or movies. In education, QA systems can be used to create educational platforms that provide students with interactive and personalized learning experiences. These systems can be used to answer students' questions in real-time and provide additional information and resources to support their learning. In business intelligence, QA systems can be used to extract important information from large datasets and provide insights to business decision-makers. These systems can be used to answer questions about sales trends, customer demographics, and other business-critical data. In healthcare, QA systems can be used to provide doctors and nurses with quick and accurate answers to medical queries. These systems can be used to answer questions about treatments, medications, and patient history, improving the quality of care and reducing the risk of errors. In research, QA systems can be used to extract information from scientific literature and assist researchers in finding relevant papers, tracking progress in specific fields, and gaining new insights.

2.2 Visual-Question-Answering

VQA has gained much interest in the research community, with many VQA datasets proposed. The first large-scale dataset on VQA was proposed by Antol et al. [3]. Early works [18, 38, 65, 73] mainly focus on natural images and the ability to understand the visual and textual contents. Recent works [26] introduce questions requiring more sophisticated visual and semantic reasoning.

Besides natural images, abstract diagrams are also common. Many VQA datasets on abstract diagrams have been proposed to fulfill the requirement of reasoning between language and diagrams. For example, NLVR [66], FigureQA [29], and DVQA [28] focus on VQA on scientific diagrams. These datasets consider various types of figure and question templates. Another line of work focus on math and science problems with more realistic and complicated scenarios [63, 34, 61, 62, 47]. Some other works include the task of providing responses to queries about conceptual illustrations like abstract VQA [3, 82]. As some datasets often require specific domain knowledge and make it difficult to separate the visual reasoning and domain knowledge, IconQA[51] was proposed as a mathematics abstract diagram VQA dataset that only regards elementary commonsense. The questions in IconQA were

inspired by math problems for children, requiring diverse reasoning skills and containing various abstract icon objects. This study focuses on the mathematics abstract diagram VQA tasks in IconQA.

Many approaches have been proposed for the VQA tasks. Early methods often combine visual and textual inputs using attention mechanisms [36, 49, 48, 17, 79, 16]. As the semantics in the visual context are often presented by specific objects, some works utilize object proposals to create the visual features [36, 79, 16]. Inspired by the good performance of Transformer models, some works employed pre-trained language-vision models for the VQA tasks in natural images and obtained noticeable improvement. As most pre-trained models are trained on natural images and are not efficient in abstract diagram VQA, Lu et al.[51] introduced the Icon645 dataset for pre-training backbone networks like ResNet [21] to obtain better visual representations from abstract diagrams. Patch-TRM[51] is a cross-modal Transformer model employing the image patch strategy and a ResNet pre-trained on Icon645 to acquire the visual representations. However, features that come from image patch tokens or object proposals may not efficiently capture and present the information about the type and number of objects, which is essential for answering the mathematical question. The description augmentation proposed in this study aims to capture the important aspects of the visual information and help answer the mathematical question.

VQA systems have a broad spectrum of utilization across multiple industries. For image captioning, VQA systems can be used to produce textual descriptions for visual content, providing a more detailed and comprehensive understanding of the image content. In robotics, VQA systems can be used to enhance the capability of robots in identifying objects, navigation, and manipulation by providing them with the ability to understand and respond to natural language commands. For surveillance, VQA systems can be used to enhance the capabilities of surveillance systems by providing them with the ability to understand and respond to natural language queries about the content of surveillance images. In human-computer interaction, VQA systems can be used to improve the performance of human-computer interaction systems by providing them with the ability to understand and respond to natural language queries about images and videos. In augmented reality, VQA systems can be used in augmented reality systems to furnish users with pertinent data about the objects in the real-world by answering natural language queries. In E-commerce, VQA systems can be used to improve e-commerce systems' performance by providing customers with the ability to search for products by asking natural language queries about images and videos of products.

The potential of VQA systems in education is significant, as it has the

ability to revolutionize the way students interact with and learn from educational materials. VQA systems, which are capable of understanding and responding to natural language queries about images and videos, can be used to create interactive and personalized learning experiences for students. One potential application of VQA in education is in the creation of interactive educational materials. These systems can be utilized to generate natural language explanations for images and videos, allowing students to interact with educational materials in a more intuitive and engaging way. This can aid in the improvement of student comprehension and retention of information. Furthermore, the ability of VQA systems to provide context-aware and personalized explanations can adapt to the diverse abilities of students, making education more inclusive. Another potential application of VQA in education is in the creation of virtual tutors. VQA systems can be used to provide students with real-time feedback and support, answering their questions and providing additional information and resources to support their learning. This can help to improve the effectiveness of online and distance learning, enabling students to receive assistance even when not in a traditional classroom setting. Additionally, VQA systems can be employed to create educational games and interactive quizzes, which can make learning more enjoyable and engaging. This can help to improve student motivation and engagement, encouraging students to take an active role in their own learning. The utilization of VQA in education is still in its infancy, but as technology continues to evolve and improve, it is expected that it will play an increasingly important role in the field of education. By providing students with interactive and personalized learning experiences, VQA systems can help to improve their understanding and retention of information, ultimately leading to better educational outcomes.

2.3 Finding information in Scientific Papers

The inability of most Transformer-based models to handle long sequences is a result of the quadratic scaling of the self-attention operation. This has led to the development of new architectures that can handle these types of sequences. For example, Beltagy et al. [4] proposed a model that utilizes an attention mechanism that scales linearly with sequence length, allowing it to process documents containing thousands of tokens. Ainslie et al. [1] presented the Extended Transformer Construction (ETC) which addresses the challenges of scaling input length and encoding structured inputs that are present in standard Transformer architectures. Zaheer et al. [80] further extended ETC to more generic scenarios where there may be no prior domain

knowledge about the structure of the source data, referred to as BigBird. However, the ability of these models to capture semantics may not be as efficient when dealing with shorter sequences due to the sheer volume of information they have to process.

Retriever-Reader is a prevalent approach in current open-domain question answering systems. A Retriever, usually considered an Information Retrieval (IR) system, is responsible for identifying and retrieving relevant documents related to the given question. The Reader, typically a neural Machine Reading Comprehension (MRC) model, then extracts the final answer from the retrieved documents.

Traditionally, DrQA employed sparse retrievers (TF-IDF, BM25) to identify relevant documents. However, recent advancements have proposed dense retrievers, which represent both text as dense vectors. There are two main strategies, shared parameters [59][54] or two distinct encoders [31][19][42][64].

In addition to retrievers, Transformers-based re-rankers have also been implemented in retrieval-based QA systems [25][52][74][55][56][77][35][15]. One popular approach for re-rankers is the Cross-encoder.

Chapter 3

Proposed Method

3.1 Description Augmentation for VQA in Mathematics Abstract Diagram

3.1.1 Problem formulation

The proposed method aims to solve the three abstract diagram VQA subtasks in IconQA [51]. The three tasks involve reasoning and basic mathematics skills from pre-kindergarten to third grade. Figure 3.1 shows examples of the dataset.

In task *multi-image-choice*, a question is given in natural language along with N candidate images ($1 \leq N \leq 5$) and a context image. The context image is optional and can be absent for some questions. The system has to choose the correct image from N images that best answer the given question. The *multi-text-choice* sub-task is similar to *multi-image-choice* sub-task but the candidate answers are texts instead of images. For each problem, the system is given a natural language question along with N text options($1 \leq N \leq 5$) and a context image. Different from the *multi-image-choice* sub-task, the context image is always given. The system has to choose the correct answer from N text options that best answers the given question. In the *fill-in-the-blank* task, the system is given a question along with a visual context and is required to output a short text-based answer for the given question.

3.1.2 Object Detection for Abstract Diagram

Intuitively, to solve abstract diagram mathematics problems like those in Figure 3.1, we first need to identify each object in the visual information. The information about the class, position and numbers of the objects is es-

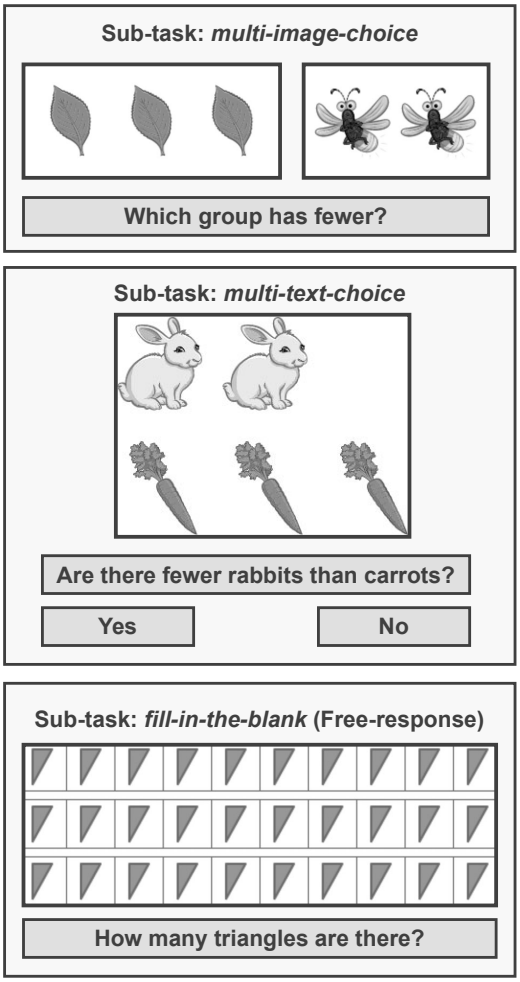


Figure 3.1: Examples of three sub-tasks in IconQA

sential in solving the problem. Therefore, the first part of the proposed system is the object detection component. This component aims to detect the objects in the provided visual information. Unlike VQA in natural images [3, 83, 18, 26, 24, 73] or abstract scenes [82] where every detail in the image is connected to each other, VQA in mathematics abstract diagrams like IconQA [51] has separate abstract icon objects. Although object detection methods [60, 20, 7] can be used to detect the objects in the visual information, the pre-trained models of these methods are trained on natural images and are not inadequate for the detection of abstract icons. Lu et al. [51] propose Icon645, an icon dataset containing 377 classes, for training icon classification models. However, to our knowledge, no dataset with annotated bounding boxes is available for fine-tuning object detection models on

abstract icons. One important property of the visual information in mathematics abstract diagram VQA is that the objects usually appear on the same simple background and are separate from each other. Leveraging this property, we propose using Connected-Component Labeling, a connected regions detecting technique for binary images, to detect the object in the visual information. Many algorithms are available for Connected Component Labeling [23, 69, 22]. For the proposed system, we implement a simple algorithm that identifies the connected components and the corresponding bounding boxes. Algorithm 1 shows how to identify all connected components of a given pixel in an image. The input of the algorithm is an image and a pixel position (row index and column index). The algorithm outputs the top-left and bottom-right coordinates of the bounding box containing all connected components. Algorithm 2 shows how to detect objects in an image using the connected components identified by Algorithm 1. The input is an image and the output is the bounding boxes of detected objects.

Algorithm 1 Identify connected components of a given pixel

Require: $image, row, column$ (position of the given pixel)
  top, $bottom \leftarrow row$
  $left$, $right \leftarrow column$
  $direction \leftarrow [(-1, 0), (1, 0), (0, -1), (0, 1)]$
  $queue \leftarrow []$
  $queue$.push($(row, column)$)
  while $queue$ **do**
    $current\_row$, $current\_column \leftarrow queue.pop()$
    for d_r, d_c in $direction$ **do**
      $new\_row \leftarrow current\_row + d\_r$
      $new\_column \leftarrow current\_column + d\_c$
      if $image[new\_row, new\_column]$ is valid **then**
        $image[new\_row, new\_column] \leftarrow$ invalid
        $top \leftarrow min(new\_row, top)$
        $bottom \leftarrow max(new\_row, bottom)$
        $left \leftarrow min(new\_column, left)$
        $right \leftarrow max(new\_column, right)$
        $queue$.push((new_row, new_column))
      end if
    end for
  end while
  return top, $left$, $bottom$, $right$

Algorithm 2 Object detection via connected-component labeling

Require: *image*
  bounding_boxes ← []
  for *row* ← 0 to *image.height* **do**
    for *col* ← 0 to *image.width* **do**
      if *image*[*row*, *col*] is valid **then**
        bbox ← *get_connect_components*(*image*, *row*, *col*)
        bounding_boxes.append(*bbox*)
      end if
    end for
  end for
  return *bounding_boxes*

3.1.3 Description Generation

Figure 3.2 shows the overall description generation process. Given an input image, the generation process will output the corresponding description. First, the Connected-Component Labeling is applied to detect separate objects in the input image. Specifically, the top-left and bottom-right coordinates of the bounding boxes of objects are returned from the Connected-Component Labeling component. Each object in the image is cropped using these coordinates and fed to a ResNet model[21]. This ResNet model is pre-trained on Icon645 dataset [51] for icon classification. After the classification step, the names of the object classes are collected for description creation. The classes with the same name are then grouped together. For each class, the description is created using the pattern *¡number of object¿ ¡class name¿*. These descriptions are then concatenated to form the final description for the input image. For example, if the collection of class names after the classification step is {*rabbit, rabbit, carrot, carrot, carrot*}, the final description will be *2 rabbits, 3 carrots*. The generated descriptions can be used to augment the information in the original problem to train vision-language VQA models. Moreover, the descriptions can also be used to train question-answering models without using visual information. Both of these approaches are shown in the experiments.

3.1.4 Description Augmentation for VQA Models

The generated descriptions can be used to augment the training and inference process of existing VQA models and enhance the performance of these models. Naturally, there are two sources of information in a training exam-

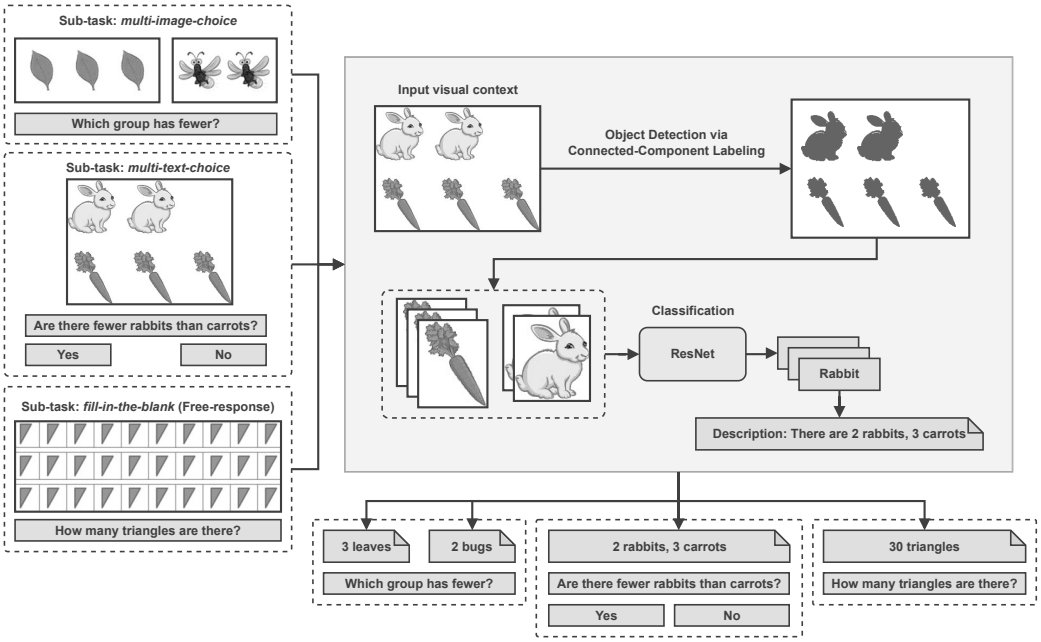
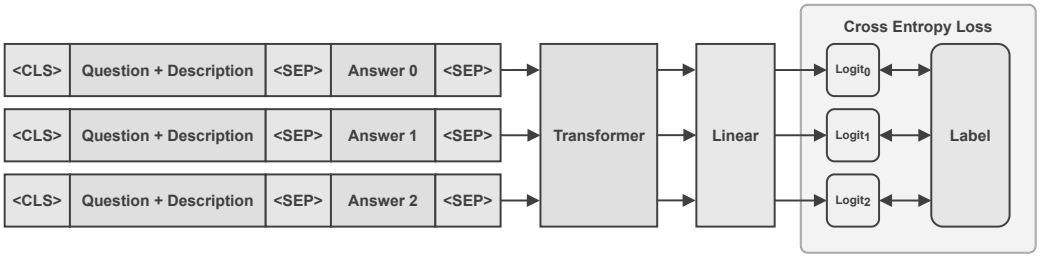


Figure 3.2: Description Generation Process

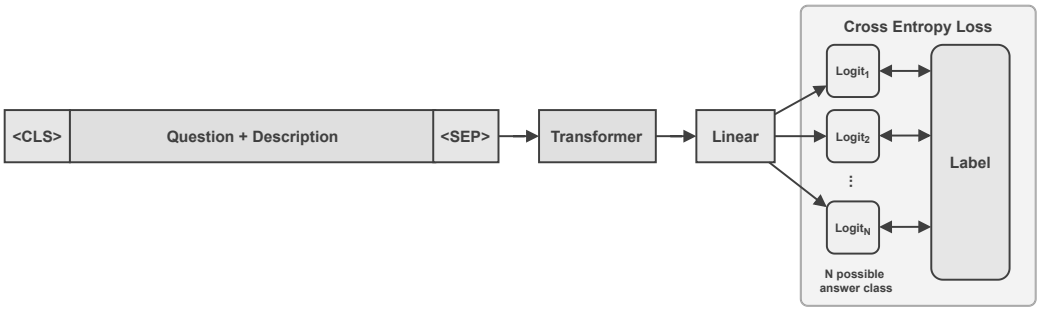
ple: the textual information from the question and the visual information from the diagram. Most VQA models process textual and visual information separately. Therefore, a straightforward way to augment the training process using the generated descriptions is to concatenate the descriptions and the questions. The resulting sequences after concatenation are used as the new questions for training. The same approach is used in the inference process; The generated description and the question are concatenated before being fed to the VQA model for answer prediction. Wallace et al. [70] discovered that many token embedding methods, including Transformers like BERT[13], can naturally encode numeracy and perform numerical reasoning to some degree. As a result, questions augmented with generated descriptions may provide helpful information for VQA in the training and inference process. The experiments show that the performance of existing VQA models can benefit significantly from the description augmentation.

3.1.5 Description-based Models

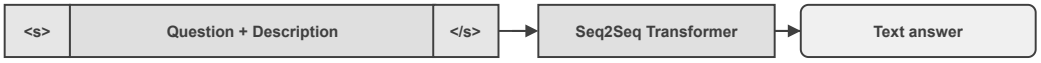
We developed different models to evaluate the effectiveness of the descriptions in answering the mathematics questions. These models only use the generated descriptions and do not consider the original visual information from the problems. We refer to these models as *description-based models*. As



(a) Textual multiple-choice question-answering model, used for *multi-image-choice* and *multi-text-choice* tasks



(b) Multi-class classification model used for *fill-in-the-blank* task by choosing one from all possible N answers from the training set.



(c) Sequence-to-sequence text generation approach for *fill-in-the-blank* task

Figure 3.3: Models for mathematics abstract diagram VQA using only the generated descriptions. No visual information is used.

the images are replaced by the generated descriptions, the tasks *multi-image-choice*, *multi-text-choice* can now be both viewed as textual multiple-choice question-answering task, and the task *fill-in-the-blank* can be viewed as conditional text generation task.

We developed a Transformer-based multiple-choice question-answering model for the *multi-image-choice* and *multi-text-choice* tasks. The model only uses the textual information from the descriptions and does not consider the visual information. The architecture of the model is depicted in Figure 3.3a, where the input comprises the question, the generated description, and a maximum of five answer options, represented by N ($1 \leq N \leq 5$). The model outputs the score for each option, and the option with the highest score is chosen as the answer. Our model follows the approach of fine-tuning BERT[13] on the SWAG dataset [81]. A linear layer is stacked on the Transformer model to work as a multiple-choice classification head. For each

example, with the condition $1 \leq N \leq 5$, we construct five input sequences by concatenating the question, the descriptions, and an answer option. The separate-token *SEP* and the classification-token *CLS* are added to the sequences. Each sequence is then tokenized and fed to the Transformer model separately. The multiple-choice classification head outputs a logit for each answer option. A *cross-entropy-loss* is then calculated using the logits and the answer label.

The task of "fill-in-the-blank" requires the model to utilize solely textual information from the provided descriptions, without taking visual information into account. Two approaches were employed in the development of models for this task. The first approach treated the problem as a multiple-class classification task, where the model outputs its response by choosing one option from the possible responses in training data. The overall architecture of the model, as shown in Figure 3.3b, is similar to the approach of Devlin et al.[13] in the GLUE task [71]. A linear layer, which acts as a sequence classification head, was added to the Transformer model and outputs a logit for each answer class. The logits and the answer label are then used to calculate the *cross-entropy-loss*. In the second approach, the problem was viewed as a conditional sequence generation task, and the sequence-to-sequence approach of BART [43] was employed. The overall architecture of this method is shown in Figure 3.3c. The input sequence is formed by combining the question and the description, which is then fed to the BART model to generate the text answer.

3.2 Finding information in scientific papers

We also propose a retriever-reader method for finding information in scientific documents. The proposed method for information retrieval in scientific documents is a combination of a retriever and a reader, with the incorporation of a sliding window technique to enhance the efficiency of the pipeline. This approach is tailored to addressing open-domain questions and involves breaking the process down into two separate stages: identifying relevant passages and extracting the final answer.

In the passage retrieval phase, a corpus comprising of an article is utilized, and a retriever is employed to identify the pertinent passages in response to a question $Q = (q_1, ..., q_{|Q|})$. The methodology adopted in this phase is in accordance with the approach outlined in the studies conducted by Karpukhin et al. in [31] and Wang et al. in [74], which involves breaking down the text into smaller, non-overlapping segments of a specific word count, referred to as passages, $P = [P_1, ..., P_i, ..., P_m]$, where $P_i = (p_i^1, p_i^2 ..., p_i^{|p_i|})$ is

the i-th passage, $P_i \in A$, $P_i \in A$, and $q_k \in Q$ and $p_i^j \in P_i$ are corresponding words and serves as the basic units of retrieval.

In the answer extraction phase, only questions that have extractive answers are considered, and a machine reading comprehension model, known as a reader, is employed to identify the specific answer within the relevant passages identified in the passage retrieval phase. Given a question $Q = (q_1, ..., q_{|Q|})$ and a context passage $C = (c_1, c_2, ..., c_n)$. The answer extraction phase utilizes a reader to identify the specific text within the relevant passages that answers the question. This process involves pinpointing a specific text span, represented as $(c_i, c_{i+1}, ...c_j)$ within the context that is determined to be the answer. The experiments carried out with this method only consider the passage that was assigned the highest relevance score during the retrieval phase. The overall approach is depicted in a diagram in Figure 3.4.

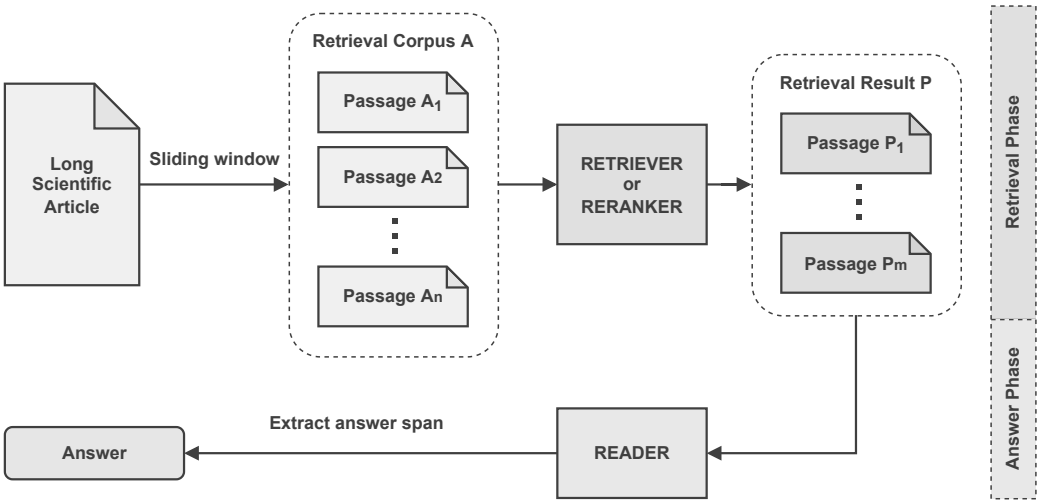


Figure 3.4: Retriever-Reader approach on long scientific articles

To preprocess articles, we divide them into disjoint text blocks of WW words, known as passages. However, this method may cause loss of context for answer spans near the boundary. We employ a sliding window technique, inspired by Wang et al. [74], with a stride of S=50 for all experiments. We also experiment with various values of WW to determine the optimal value.

For the retriever component, we investigate two types of approaches: retrieving and re-ranking. The retrieval process employs a TF-IDF indexing system to identify the passages within an article that are most relevant to a given question. This is done by calculating the cosine similarity between the question and each passage, which allows the system to identify the top

most relevant passages. The re-ranking approach, on the other hand, utilizes a Transformer cross-encoder model to generate a score between 0 and 1 for each question-passage pair, indicating the level of relevance of the passage to the question.

The reader component of our system utilizes Transformer-based models. The retriever-reader approach allows us to use any cutting-edge Transformer-based model for QA, without the limitation of handling long sequences. As mentioned earlier, the goal of the reader is to identify a specific text segment $(c_i, c_{i+1}, ...c_j)$ from $C = (c_1, c_2, ..., c_n)$ that answers the question. Our approach for extracting the answer to a question utilizes a pre-trained Transformer model, following the methodology outlined by Wolf et al. in their work [76]. This includes adding a linear layer to the model's output, which allows us to calculate the likelihood of a given token being the start or end of the text span that answers the question. Specifically, we calculate span start probabilities $Pstart_i$ and span end probabilities $Pend_i$ for c_i. For each span $(c_i, c_{i+1}, ...c_j)$, a score is calculated using $Pstart_i$ and $Pend_j$; The text segment with the maximum score is chosen as the response to the inquiry.

Chapter 4

Experimentation and Evaluation

4.1 Description Augmentation for VQA in Mathematics Abstract Diagram

4.1.1 Experimental Setup

Method	Sub-tasks (3)			Reasoning skills (13)												
	Img.	Txt.	Blank	Geo.	Cou.	Com.	Spa.	Sce.	Pat.	Tim.	Fra.	Est.	Alg.	Mea.	Sen.	Pro.
Human	95.69	93.91	93.56	94.63	97.63	94.41	93.31	92.73	95.66	97.94	97.45	87.51	96.29	86.55	97.06	85.67
Top-Down[2]	75.92	68.51	73.03	80.07	65.01	80.65	45.78	58.22	55.01	68.28	72.43	99.54	50.00	99.46	84.54	83.75
BAN[36]	76.33	70.82	75.54	79.99	67.56	82.12	53.20	66.92	55.67	66.50	73.77	97.06	47.46	96.50	82.12	82.45
ViLBERT[46]	76.66	70.47	77.08	80.05	71.05	75.60	49.46	58.52	62.78	66.72	74.09	99.22	50.62	99.07	81.78	70.94
MCAN[79]	77.36	71.25	74.52	79.86	68.94	82.73	49.70	62.49	54.79	68.00	76.20	99.08	47.32	98.99	83.25	84.87
DFAF[16]	77.72	72.17	78.28	81.80	70.68	81.69	51.42	67.01	56.60	67.72	77.60	99.02	50.27	98.83	84.11	85.70
UNITER[9]	78.71	72.39	78.53	81.31	71.01	83.67	48.34	61.25	60.81	69.77	78.37	99.41	49.18	99.38	86.10	87.84
ViT[14]	79.15	72.34	78.92	82.60	70.84	82.12	54.64	68.80	58.46	68.66	77.41	98.95	51.10	98.76	84.72	86.07
ViLT[37]	79.67	72.69	79.27	82.61	71.13	84.95	53.38	66.72	59.22	69.99	75.81	99.02	50.55	98.91	86.10	87.65
Patch-TRM[51] (Reproduced)	79.04	67.37	79.47	78.65	74.59	74.09	53.52	58.27	54.46	75.49	81.49	97.19	53.23	97.13	91.28	75.02
Patch-TRM[51] (From paper)	*82.66*	*75.19*	*83.62*	*81.87*	*77.81*	*87.00*	*55.62*	*62.39*	*68.75*	*77.98*	*82.13*	*98.24*	*56.73*	*97.98*	*92.49*	*95.73*
Models using description augmentation																
Patch-TRM (Desc augment) (1)	**80.74**	70.74	**84.14**	82.32	75.82	**78.33**	53.94	62.05	61.19	**75.87**	**81.49**	**98.69**	55.01	**98.45**	**92.57**	**75.86**
Description-based Model (2)	78.72	**78.23**	68.64	**85.99**	**83.50**	75.67	**67.06**	**76.65**	**63.33**	67.83	61.77	67.84	**73.21**	61.77	86.79	74.09

Table 4.1: Accuracy on IconQA test set. Patch-TRM (Reproduced) is from reproducing experiments using the provided code, used for comparison. Patch-TRM (From paper) is taken from the cited paper, used for reference. The proposed description augmentation method is evaluated in two settings: (1) Description augmentation for existing VQA models. (2) Description-based models using only the descriptions without visual information.

Model	Training data	Sub-tasks		
		multi-image-choice	multi-text-choice	fill-in-the-blank
RoBERTa-Large	multi-image-choice	78.72	51.88	-
RoBERTa-Base	multi-image-choice	78.63	48.37	-
BERT-Base	multi-image-choice	78.61	44.25	-
RoBERTa-Large	multi-text-choice	39.61	78.23	-
RoBERTa-Base	multi-text-choice	41.23	77.23	-
BERT-Base	multi-text-choice	43.71	78.02	-
RoBERTa-Large	multi-image-choice + multi-text-choice	78.51	78.29	-
RoBERTa-Base	multi-image-choice + multi-text-choice	79.01	77.87	-
BERT-Base	multi-image-choice + multi-text-choice	78.13	78.28	-
BART-Large	fill-in-the-blank	-	-	68.64
BART-Base	fill-in-the-blank	-	-	68.61
RoBERTa-Large	fill-in-the-blank	-	-	68.69
RoBERTa-Base	fill-in-the-blank	-	-	67.29
BERT-Base	fill-in-the-blank	-	-	66.36

Table 4.2: Accuracy on IconQA test set of different description-based models

Benchmarks and Baselines

All of the experiments are conducted on the IconQA dataset [51]. IconQA contains three sub-tasks as described in Section 3.1.1. Table 4.4 shows the overall information of IconQA. The questions in IconQA require different cognitive and mathematics skills that can be categorized into 13 groups listed in Table 4.3. We compare the proposed method with strong baselines in attention-based and Transformer-based vision-language approaches. Specifically, for attention-based methods, the baselines include Top-down attention [2], BAN [36], DFAF [16], and MCAN [79]. For Transformer-based vision-language approach, the baselines include VilBERT[46], UNITER [9], ViT [14], and ViLT [37]. The performance of these attention-based and Transformer-based methods is taken from the work of Lu et al. [51] for comparison. The proposed method is also compared with Patch-TRM[51], a cross-modal Transformer model employing a ResNet pre-trained on Icon645[51]. We reproduced the results of Patch-TRM from the provided code from the work of Lu et al. [51] and used the reproduced results for comparison. We also conducted experiments on Patch-TRM trained with description augmentation to evaluate the impacts of the augmented descriptions on the performance of existing VQA models.

Implementation Details

The proposed components, including the Connected-Component Labeling object detection and description-based models, are implemented using Python and PyTorch. The description-based multiple-choice (for *multi-image-choice*,

multi-text-choice tasks) and multi-class classification model (for *fill-in-the-blank* task) used RoBERTa-Base, RoBERTa-Large[45], BERT[13] as its Transformer component. The description-based sequence-to-sequence text generation (for *fill-in-the-blank* task) used BART-Large, BART-Base [43] as its Transformer component. The implementation of Transformer models follows the work of Wolf et al. [75]. The code and models are available.

Skill types	Description
Geometry	Recognize shapes, symmetry, transformations
Counting	Identify the number of shapes, objects
Comparing	Compare attributes of objects
Spatial	Identify spatial positions and relations
Scene	Comprehend abstract scenes
Pattern	Recognize different patterns
Time	Recognize clock time, event time
Fraction	Understand fractions
Estimation	Estimate big numbers, sizes
Algebra	Conduct algebraic operations
Measurement	Measure lengths, widths, heights
Commonsense	Contain external knowledge
Probability	Fulfill operations of probability/statistics

Table 4.3: Reasoning skill categories in IconQA

Tasks	All	Train	Val	Test
Multi-image-choice	57,672	34,603	11,535	11,535
Multi-text-choice	31,578	18,946	6,316	6,316
Filling-in-the-blank	18,189	10,913	3,638	3,638
All	107,440	64,462	21,489	21,489

Table 4.4: Overall of IconQA dataset

4.1.2 Results

Table 4.1 shows the performance of the baselines and the proposed method on the IconQA test set. The proposed description augmentation method is evaluated in two settings. In the first setting, the generated descriptions are used to augment the training and inference process of the existing VQA

model Patch-TRM [51], as described in Section 3.1.4. In the second setting, we evaluated description-based models using only the generated descriptions without any visual information, as described in Section 3.1.5. The result suggests that the performance of Patch-TRM can benefit noticeably from the description augmentation. Moreover, the result indicates that the proposed description-based models achieved competitive results in *multi-image-choice* sub-tasks and state-of-the-art results in *multi-text-choice* sub-tasks, using only the signals from the descriptions.

To gain a deeper understanding of the impact that the generated descriptions have on different reasoning skills, we visualize the accuracy of the models on 13 skill categories in Figure 4.1. Figure 4.1a compares the Patch-TRM model [51] and our description-based model described in Section 3.1.5. While the Patch-TRM used both the textual and visual information from the problem, the description-based model only considered the textual information from the descriptions. The result suggests that while the Patch-TRM shows strong performance on skills that require recognizing visual properties of objects (*Commonsense*, *Measurement*, *Estimation*, *Fraction*), the description-based model outperforms in mathematics skills that require knowing the properties and quantities of objects (*Counting*, *Comparing*, *Algebra*, *Pattern*). Surprisingly, the description-based model also outperforms Patch-TRM in the *Spatial* skill which requires identifying spatial positions and relations.

We experiment on different aspects of the proposed description-based models and report the results in Table 4.2. As we only consider the textual descriptions, the *multi-image-choice* and *multi-text-choice* sub-tasks can both be seen as multiple-choice question-answering and can be used to jointly train a model. Although it is evident in the result that models trained on one sub-task can have some knowledge of the other sub-task, the result indicates no significant difference between jointly training the two sub-tasks and training them separately. Moreover, the two approaches for *fill-in-the-blank* task, using multi-class classification and sequence-to-sequence text generation, witness approximately the same result in performance. The result also shows that the choice of Transformer models (RoBERTa-Large, RoBERTa-Base, BERT-Base) has little impact on the performance.

We analyze some examples of the generated descriptions to gain a deeper understanding of the generation process. We randomly sample 100 examples and report the three most common cases in Figure 4.2. Figure 4.2a shows an example where the description is correct and provides useful information to answer the question. This example is the type of descriptions we aim to generate. Figure 4.2b shows a flawed example where the number of objects is correct, but the object class is wrong. In this example, the proposed object

detection method works correctly, but the ResNet classification model fails to classify the correct class. These flawed examples still provide some helpful information and can be improved by enhancing the classification model. Figure 4.2c shows some bad examples where the descriptions do not contain any useful information to answer the questions. This problem often comes from overlapping objects in the image or the nature of the question requiring a more sophisticated description. There are fruitful areas for future research regarding this problem.

4.2 Finding information in Scientific Papers

4.2.1 Experimental setup

Dataset

In this research, we evaluate our method using the QASPER dataset[11] which comprises of a total of 5049 questions and 1585 NLP papers. The QASPER questions were created by individuals who only read the title and abstracts of the papers, and a separate group of individuals provided the answers. The dataset is divided into four categories: extractive, abstractive, yes/no, and unanswerable. Our study concentrates on the extractive category which accounts for 51.8% of the entire dataset.

Baselines

In this study, we evaluate the performance of our proposed method against other state-of-the-art models in the QASPER dataset [11]. Specifically, we compare our results to those obtained by DocHopper [67] and QASPER-LED [11]. DocHopper is an iterative model that navigates through different sections of hierarchical documents to answer complicated queries. An alternate approach we examine is QASPER-LED, a model constructed using the Longformer architecture [4], which can handle input sequences of up to 16,000 tokens. Additionally, we compare our method against the ETC reader [1], a highly-regarded model for handling long sequences with a maximum input length of 4096 tokens. The performance results for the ETC reader on the QASPER dataset are sourced from the research of Sun et al. [67] and were evaluated using two distinct techniques: sequential reading and retrieval with a BM25 retriever. The official QASPER evaluation metric, F1 score, is used to assess performance. The F1 score takes into consideration the token-level overlap between the predicted and gold answers, and is calculated using the following formula.

$$\text{Precision} = \frac{\text{Count of correct token predictions}}{\text{Count of token predictions}}$$

$$\text{Recall} = \frac{\text{Count of correct token predictions}}{\text{Count of tokens in gold answer}}$$

$$\text{F1} = \frac{2\ .\ \text{Precision}\ .\ \text{Recall}}{\text{Precision} + \text{Recall}}$$

Implementation details

In the course of our experimentation, we selected TF-IDF as a sparse retrieval method and Transformer-based Cross-encoder as a re-ranking approach for comparative analysis. The implementation of the TF-IDF retriever was carried out using the Scikit-learn library[57]. The Cross-encoder re-ranking model was implemented utilizing the SBERT[59] and Transformers[76] pre-trained model built on MiniLM[72] and fine-tuned on the MSMARCO[6] dataset. The readers were constructed using the Transformers library[76] and pre-trained models RoBERTa[44] and ELECTRA[10]. In certain experiments, we also introduced the *Oracle* setting as a performance upper bound for the retrievers. The *Oracle* setting involves the use of passages containing the gold answers as system input rather than the entire article, taken from the annotated evidence in QASPER.

4.2.2 Experimental Results

Method analysis

A series of tests were conducted on the QASPER development dataset to evaluate various aspects of the proposed method. One set of experiments looked at how different window sizes impacted the retrieval phase and overall QA pipeline. The results, shown in Figure 4.3, demonstrate how varying the window size and number of retrieved passages affects retrieval accuracy, which is measured as the percentage of retrieval results that include at least one correct answer. Another set of experiments, illustrated in Figure 4.4, shows the performance of RoBERTa and ELECTRA readers, using large versions of both models. The highest performance for RoBERTa and ELECTRA readers was seen with a window size of 150, resulting in F1 scores of 28.66% and 30.43%, respectively. The pipeline's performance was found to be higher when using the sliding window technique, as seen in both readers at all window sizes.

In our second series of experiments, we evaluated the effectiveness of three distinct methods of retrieval, specifically the Sparse Retriever, Dense Retriever, and Cross-encoder Re-ranker. To gauge the impact of each approach on the overall QA performance, we plotted the results in Figure 4.5 using the ELECTRA reader for all the experiments. As previously mentioned in the methodology section, the "Oracle" represents the highest achievable performance for the retrievers. Our findings indicate that the Re-ranker consistently outperforms the Sparse Retriever and Dense Retriever across all window sizes, with a significant gap between the Re-ranker and the Oracle, suggesting room for further improvement in the retrieval stage. Additionally, we compared the retrieval accuracy of the three retrievers at different numbers of retrieved passages in Figure 4.6 and found that the Re-ranker continued to perform better than the other two retrievers, with accuracy increasing as the number of passages increased.

Comparison with competitive methods

We evaluate the performance of our method against state-of-the-art models on the QASPER development set, as presented in Table 4.5. Our approach demonstrates superior performance in comparison to other state-of-the-art models, such as ETC[1], QASPER-LED[11], and DocHopper[67], which are specifically designed to handle long sequences. These results illustrate the effectiveness of our retriever-reader approach and its ability to effectively utilize high-performance Transformer-based models such as RoBERTa, ELECTRA, regardless of sequence length.

Method	**F1 Score**
Retrieval + ETC	18.70
Sequential (ETC)	24.60
QASPER-LED	26.10
DocHopper	29.60
Proposed method - MiniLM Re-ranker - ELECTRA Reader - Sliding window (W=150)	**30.43**

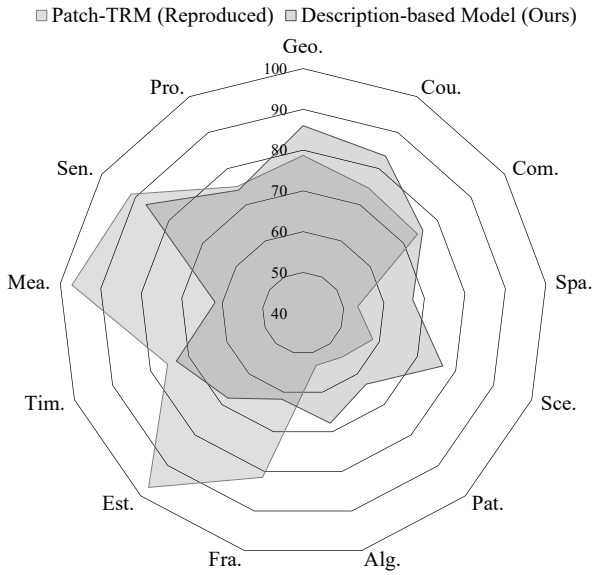
Table 4.5: Assessment of the proposed method against current state-of-the-art techniques on the QASPER development set, specifically the extractive question subset

Retriever	Reader	Sliding Windows	Similarity Score	F1 Score
Re-ranker (MiniLM)	RoBERTa	✓	-	28.66
	ELECTRA	✓	-	30.43
Sparse Retriever (TF-IDF)	RoBERTa	✓	Cosine	19.95
	ELECTRA	✓	Cosine	20.03
Dense Retriever (MPNet)	RoBERTa	✓	Cosine	18.78
	ELECTRA	✓	Cosine	19.16
Without sliding window				
Re-ranker (MiniLM)	RoBERTa	-	-	27.88
	ELECTRA	-	-	28.43
Sparse Retriever (TF-IDF)	RoBERTa	-	Cosine	18.34
	ELECTRA	-	Cosine	18.89
Dense Retriever (MPNet)	RoBERTa	-	Cosine	18.45
	ELECTRA	-	Cosine	18.53
Dot product similarity score				
Sparse Retriever (TF-IDF)	ELECTRA	✓	Dot product	20.03
	RoBERTa	✓	Dot product	19.95
Dense Retriever (MPNet)	RoBERTa	✓	Dot product	18.78
	ELECTRA	✓	Dot product	19.16
Dual-encoder				
DPR[31]	ELECTRA	✓	Dot product	20.13
	ELECTRA	✓	Cosine	18.86
Oracle				
Re-ranker (MiniLM)	ELECTRA	✓	-	54.85
Sparse Retriever (TF-IDF)	ELECTRA	✓	Cosine	54.07
Dense Retriever (MPNet)	ELECTRA	✓	Cosine	53.29

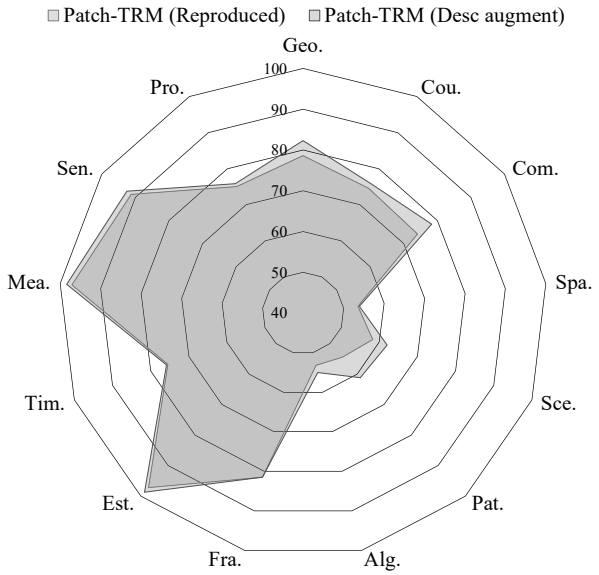
Table 4.6: An examination of individual components through removal and substitution

Ablation study

An evaluation was conducted to determine the significance of various features of the proposed method. Four specific areas were examined: the impact of the sliding window technique, the use of similarity scores in the Sparse Retriever and Dense Retriever, the effect of different architectures in the Dense Retrievers, and the upper-bound performance of the retrievers. The results of this study are summarized in Table 4.6. The first set of results compares the standard method. The second set of experiments looked at the performance without the sliding window technique, and it was determined that using a sliding window resulted in improved performance. In the third set of experiments, the dot product score was used instead of the cosine similarity score to compare the question and passage vectors in the Sparse Retriever and Dense Retriever. It was found that the results were similar regardless of which similarity score was used. The fourth set of experiments evaluated DPR[31], which is a dual-encoder architecture that uses two distinct parameterized encoders as opposed to shared parameters in the standard setting. The results showed that DPR is comparable to MPNet Dense Retriever and that the dot product score is more effective than the cosine similarity score in DPR. The final set of results represents the upper-bound performance of the retrievers, known as the *Oracle* setting.

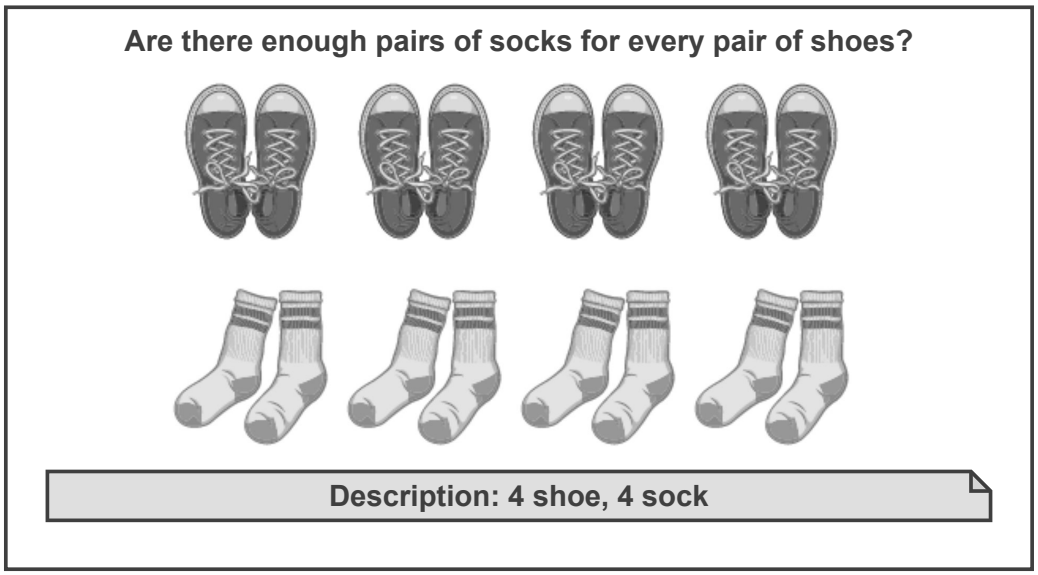


(a) Comparison between Patch-TRM and Description-based Transformer. Description-based Transformer only considers the information from the descriptions (no visual information used).

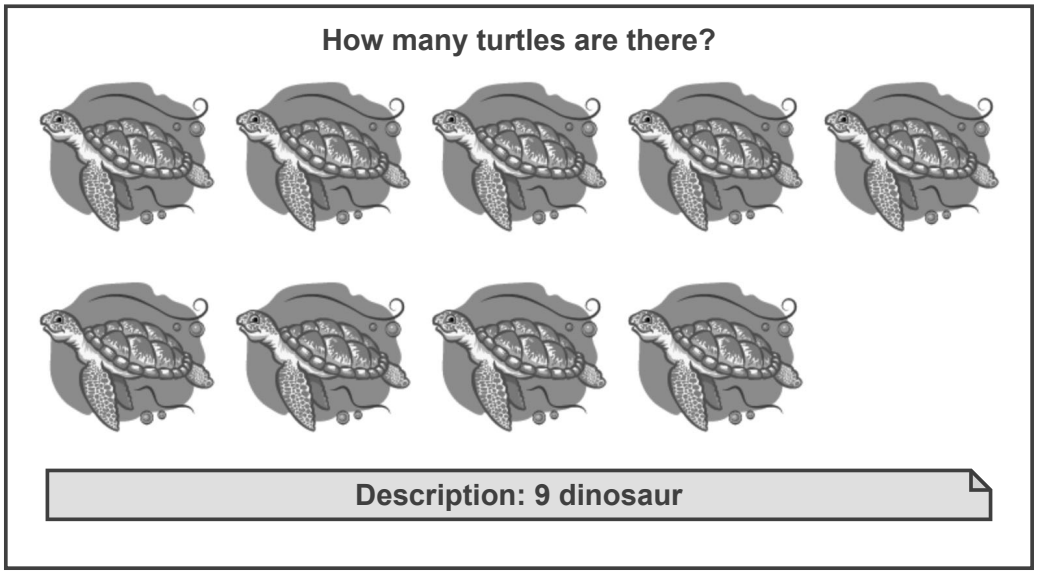


(b) Comparison between Patch-TRM model with and without description augmentation

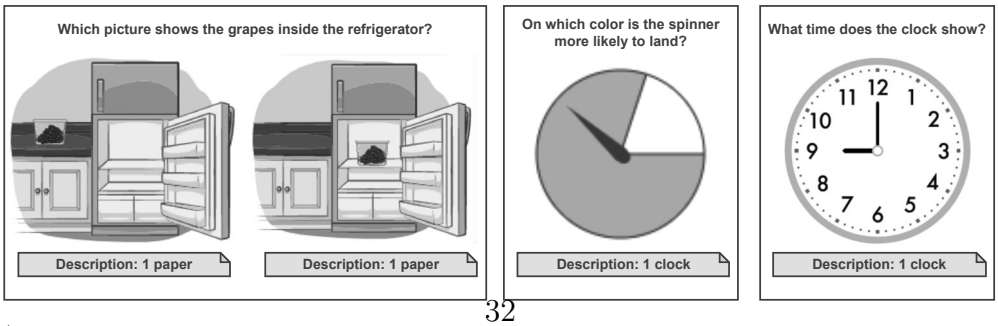
Figure 4.1: Accuracy on 13 skill categories in IconQA test set



(a) Good example where the description is correct and provides useful information to answer the question



(b) Flawed example where the number of objects is correct but the class is wrong, still provides some information



(c) Bad examples where the descriptions do not contain any useful information to answer the questions

Figure 4.2: Examples of description generation on IconQA dataset

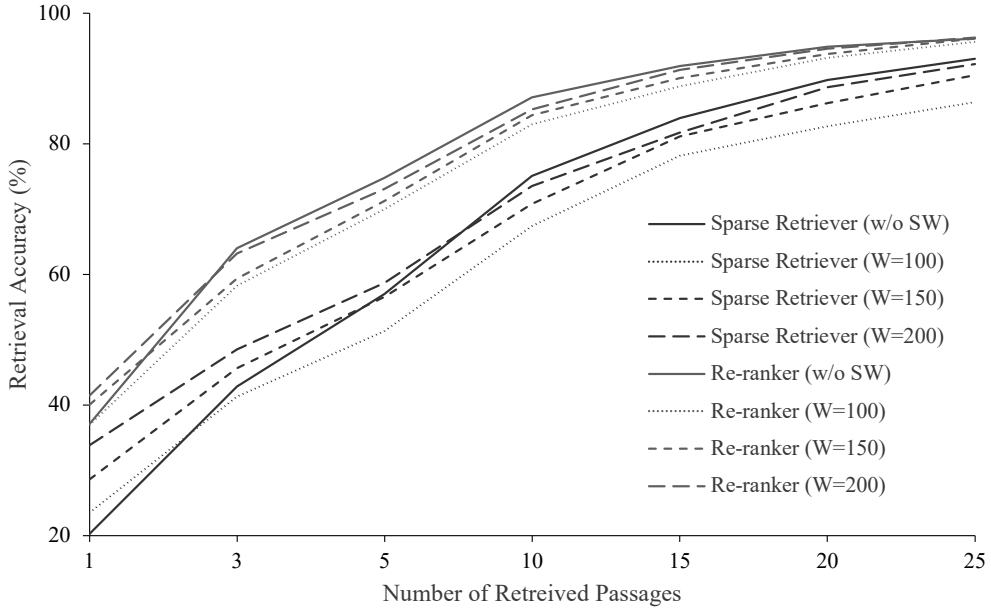
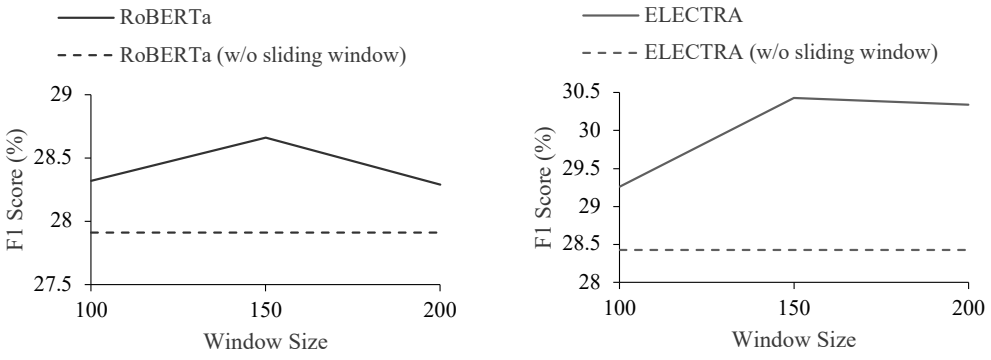


Figure 4.3: Effect of various number of retrieved passages and window size on the accuracy of Retrieval. Contrasting the performance of the Sparse Retriever utilizing TF-IDF and the Cross-encoder (CE) with diverse window size $W \in \{100, 150, 200\}$



(a) RoBERTa reader

(b) ELECTRA reader

Figure 4.4: Investigation of the impact of utilizing a sliding window technique on the performance of end-to-end question answering with various reader models

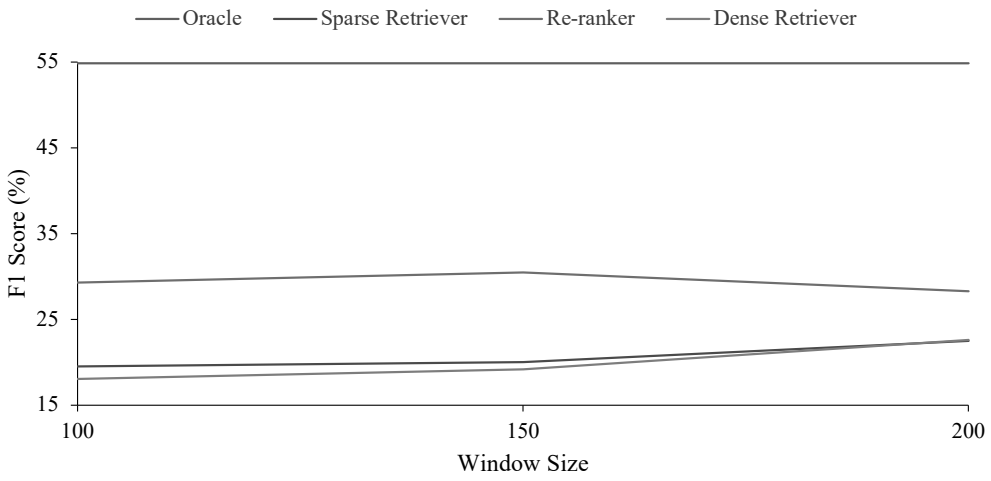


Figure 4.5: Examination of the effects of various window sizes on the performance of different retriever methods

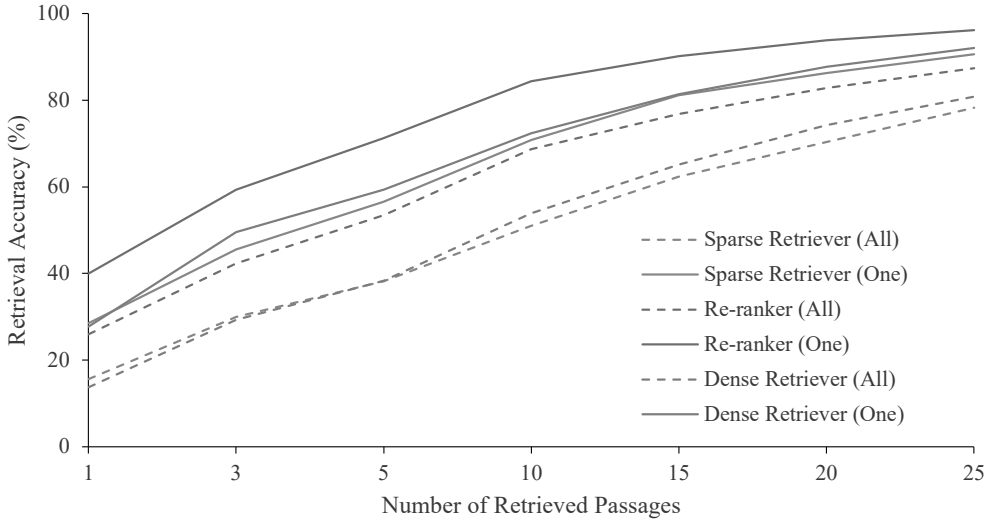


Figure 4.6: Evaluation of retrieval techniques at varying levels of retrieved passages. The solid lines indicate accuracy when at least one of the retrieved passages contains the correct answer (One). The dashed lines indicate the accuracy when all retrieved passages contain the correct answer (All)

Chapter 5

Conclusions

5.1 Conclusions

The primary objective of this research is to address the challenges associated with Visual Question Answering (VQA) systems in the domain of mathematics abstract diagrams found in textbooks and to propose an effective method for finding information in these documents. To achieve this goal, we have proposed an efficient method for generating helpful descriptions that can be utilized to augment the training and inference process of VQA models in mathematics abstract diagrams. This method aims to improve the performance of VQA systems by providing them with additional contextual information about the image. Furthermore, we have proposed an efficient method for detecting abstract objects using connected-component labeling. To evaluate the effectiveness of the proposed method, we have developed description-based models that utilize only the generated descriptions without any visual information. The experimental results demonstrate that models trained only on the descriptions can achieve state-of-the-art results in the IconQA *multi-text-choice* subtask and outperform current VQA systems on skills requiring knowledge of the properties and quantities of objects. Furthermore, VQA systems trained and inferred with augmented descriptions show substantial improvements in mathematics abstract diagram problems.

In addition, our method tackles the task of answering questions in long scientific documents by adjusting the retriever-reader approach used in open-domain QA. We employ efficient transformer-based readers to overcome obstacles presented by lengthy sequences. Additionally, we implement a technique that partitions scientific articles into fixed-size portions, which improves the performance of the retriever-reader pipeline. Our study shows that the Cross-encoder Re-rankers are more effective than Sparse Retrievers

when working with scientific articles and that a window size of 150 words provides optimal results. The QASPER dataset's experiments indicate that our approach exceeds existing leading models. This study paves the way for future research on developing efficient retrievers and readers for QA on long scientific documents, with no limitations in processing extensive sequences.

5.2 Publications

5.2.1 Publications related to the thesis

- Dieu-Hien Nguyen, **Nguyen-Khang Le**, and Minh Le Nguyen (2022). "Exploring Retriever-Reader Approaches in Question-Answering on Scientific Documents". *In: Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2022. Communications in Computer and Information Science, vol 1716.*

- **Nguyen-Khang Le**, Dieu-Hien Nguyen, and Minh Le Nguyen. "An Effective Description Augmentation Approach for Visual Question Answering in Mathematics Abstract Diagram". *(Submitted to The 32nd International Joint Conference on Artificial Intelligence (IJCAI) 2023)*

5.2.2 Other publications

- **Nguyen-Khang Le**, Dieu-Hien Nguyen, Tung Le Thanh, and Minh Le Nguyen. "VIMQA: A Vietnamese Dataset for Advanced Reasoning and Explainable Multi-hop Question Answering". *In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 6521–6529.* 2022.

- **Nguyen-Khang Le**, Dieu-Hien Nguyen, Thi-Thu-Trang Nguyen, Minh Phuong Nguyen, Tung Le, and Minh Le Nguyen. "A Novel Pipeline to Enhance Question-Answering Model by Identifying Relevant Information". *In: SCIDOCA 2021 post-proceedings (Accepted)*

- Chau Nguyen, **Nguyen-Khang Le**, Dieu-Hien Nguyen, Minh Phuong Nguyen, and Minh Le Nguyen (2022). "A Legal Information Retrieval System for Statute Law". *In: Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2022. Communications in Computer and Information Science, vol 1716.*

- Chau Nguyen, Minh-Quan Bui, Dinh-Truong Do, **Nguyen-Khang Le**, Dieu-Hien Nguyen, Thu-Trang Nguyen, Ha-Thanh Nguyen, Vu

Tran, Le-Minh Nguyen, Ngoc-Cam Le, Thi-Thuy Le, Minh-Phuong Nguyen,Tran-Binh Dang, Truong-Son Nguyen, Viet-Anh Phan, Thi-Hai-Yen Vuong, Minh-Tien Nguyen, Tung Le, and Tien-Huy Nguyen, "ALQAC 2022: A Summary of the Competition". *In: 2022 14th International Conference on Knowledge and Systems Engineering (KSE).* 2022, pp. 1-5.

- Quan Minh Bui, Chau Nguyen, Dinh-Truong Do, **Nguyen-Khang Le**, Dieu-Hien Nguyen, Thi-Thu-Trang Nguyen, Minh-Phuong Nguyen, and Minh Le Nguyen. "JNLP team: Deep Learning Approaches for Tackling Long and Ambiguous Legal Documents in COLIEE 2022". *In: JURISIN 2022 post-proceedings (LNAI) (Accepted)*

- Dieu-Hien Nguyen, **Nguyen-Khang Le**, and Minh Le Nguyen. "ViWiQA: Efficient End-to-end Vietnamese Wikipedia-based Open-domain Question-Answering Systems for Single-hop and Multi-hop Questions". *(Submitted to Information Processing & Management Journal)*

Bibliography

[1] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online, November 2020. Association for Computational Linguistics.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, 2020.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[6] Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, 2016.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[8] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, 2017.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[10] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

[11] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June 2021. Association for Computational Linguistics.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[15] Luyu Gao, Zhuyun Dai, and Jamie Callan. Modularized transfomer-based ranking framework. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4180–4190, Online, November 2020. Association for Computational Linguistics.

[16] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6632–6641, 2019.

[17] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven C. H. Hoi, and Xiaogang Wang. Question-guided hybrid convolution for visual question answering. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 485–501, Cham, 2018. Springer International Publishing.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909, 2020.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[22] Lifeng He, Xiwei Ren, Qihang Gao, Xiao Zhao, Bin Yao, and Yuyan Chao. The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognition*, 70, 04 2017.

[23] J. Hoshen and Raoul Kopelman. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Phys. Rev. B*, 14:3438, 10 1976.

[24] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019.

[25] Srinivasan Iyer, Sewon Min, Yashar Mehdad, and Wen-tau Yih. RECONSIDER: re-ranking using span-focused cross-attention for open domain question answering. *CoRR*, 2020.

[26] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017.

[27] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[28] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.

[29] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *ArXiv*, abs/1710.07300, 2018.

[30] Sevilay Karamustafaoğlu. Improving the science process skills ability of science student teachers using i diagrams. *International Journal of Physics and Chemistry Education*, 3(1):26–38, Feb. 2011.

[31] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage

retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[32] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 235–251, Cham, 2016. Springer International Publishing.

[33] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017.

[34] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017.

[35] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. *CoRR*, 2020.

[36] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[37] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021.

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, may 2017.

[39] Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi. Semantic parsing to probabilistic programs for situated question answering. In *EMNLP*, 2016.

[40] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[41] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[42] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics.

[43] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[44] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

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

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[47] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem

solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[48] Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. R-vqa: Learning visual relation facts with semantic attention for visual question answering. In *SIGKDD 2018*, 2018.

[49] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *AAAI 2018*, pages 7218–7225, 2018.

[50] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

[51] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.

[52] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Reader-guided passage reranking for open-domain question answering. *CoRR*, 2021.

[53] Maria Martiniello. Language and the performance of english-language learners in math word problems. *Harvard Educational Review*, 78:333–368, 2008.

[54] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, 2021.

[55] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *CoRR*, 2019.

[56] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *CoRR*, 2019.

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[58] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

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

[60] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[61] Mrinmaya Sachan, Kumar Dubey, and Eric Xing. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 773–784, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[62] Mrinmaya Sachan, Kumar Avinava Dubey, Tom M Mitchell, Dan Roth, and Eric P Xing. Learning pipelines with limited data and domain knowledge: A study in parsing physics problems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[63] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[64] Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. Real-time open-domain question an-

swering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441, Florence, Italy, July 2019. Association for Computational Linguistics.

[65] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[66] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[67] Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. Iterative hierarchical attention for answering complex questions over long documents, 2021.

[68] Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[69] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

[70] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. In *Empirical Methods in Natural Language Processing*, 2019.

[71] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[72] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[73] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10123–10132, 2020.

[74] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. In *EMNLP*, 2019.

[75] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[76] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[77] Ming Yan, Chenliang Li, Chen Wu, Bin Bi, Wei Wang, Jiangnan Xia, and Luo Si. IDST at TREC 2019 deep learning track: Deep cascade ranking with generation-based document expansion and pre-trained language modeling. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, volume 1250 of

NIST Special Publication. National Institute of Standards and Technology (NIST), 2019.

[78] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[79] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019.

[80] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33, 2020.

[81] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[82] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[83] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.