Exam2: Exploring the Multimodal Multilingual Assessment of Language Models

Anonymous ACL submission

Abstract

We propose Exams - a new benchmark dataset multi-lingual visual question answering for high school examinations. The special characteristic of this dataset is that the questions are part of the image. This requires the question-answering model or pipeline to first parse both the text in the input image and the diagram or picture in the image to answer the question. We are able to collect around 16,000 high-quality data samples in 10 languages covering different language families. The difficulty level of our questions are high-school level and subject ranging from physics, chemistry, mathematics, biology, politics are covered.

1 Introduction

Visual Question Answering has been considered an important task by both computer vision and machine learning communities. There have been efforts to develop datasets and model that combines computer vision, Natural Language Processing, and Knowledge Representation and Reasoning. Current datasets like ScienceQA, and Visual QA have textual questions and images and model is required to interpret the information present in both the image and the textual question and provide the correct answer. Apart from that Visual QA questions are very not very challenging with simple reasoning questions. With the advancement in both Language Models and Visual Language Models, these dataset has become very trivial. Apart from this, most of these dataset are based on English language. To address this issues, we are proposing a visual question-answering dataset. In this dataset, we a introducing a new aspect to visual question answering itself. Until now all visual question-answering dataset, each sample comprises two separate components: a textual question and an image. This setting is very artificial. In a realistic setting, the question is part of the image itself like a snapshot of a question from a textbook or exam paper. The model requires to parse both the text including complex mathematics, chemistry symbols and diagrams in the image to arrive at the correct answer. We address this setting in our dataset. Apart from this, our dataset is a multilingual dataset that covers 10 languages from diverse language families. The contributions of our work are as follows:

- In this work, we are introducing a new dimensionality to visual question answering itself. We are proposing a visual question answering task where the models need to parse both the textual question and diagram from the input image itself and do the necessary reasoning to arrive to the correct answer.
- We have proposed a multilingual dataset from the corresponding task.
- We evaluated our dataset on a large variety of existing Visual Question Answering baselines and reported our analysis.
- We also do sufficient analysis to understand both the multilingual and multi-modal capabilities of existing baselines.

2 Related Work

Question Answering has been a major focus of the NLP community for a long time with a significant focus on visual question answering in recent times. There have been efforts to create several datasets that raise different questions or aspects of visual question answering.

Science QA. ScienceQA is one most popular datasets in visual question answering. Science QA contains 21k multimodal multiple choice questions with rich domain diversity across 3 subjects (including natural science, social science, and language science), 26 topics, 127 categories, and 379 skills. The benchmark dataset is split into training, validation, and test splits with 12726, 4241, and 4241 examples, respectively. The questions are collected from elementary and high school science curricula. To answer science questions, a model needs to understand multimodal contents and also extract external knowledge to arrive at the correct answer.

Textbook QA The Textbook QA dataset raises a new question in the context of visual question answering. It proposes the task of Multi-Modal Machine Comprehension (M3C), an extension of the traditional textual machine comprehension to multi-modal data. In this paradigm, the task is to read a multi-modal context along with a multi-modal question and provide an answer, which may also be multimodal in nature. This is in contrast with the conventional question-answering task,in which the context is usually about a single modality. The TQA dataset consists of 1,076 lessons with 26,260 multi-modal questions.

MATHVISTA MATHVISTA is a consolidated Mathematical reasoning benchmark within Visual contexts. This dataset aims to encompass a diverse array of visual contexts, including natural images, geometry diagrams, abstract scenes, and synthetic scenes, as well as various figures, charts, and plots. MATHVISTA consists of 6,141 examples.

3 Exam Dataset

We introduce Exams, a new benchmark dataset for multimodal and multilingual question answering from high school examinations. The dataset has 16k samples of multi-choice questions covering different subjects. An example of Exams is shown in Figure 1. Given the image sample, the model has to first parse both the textual question and the diagram or visual context from the input image sample and select the correct answer from the multiple options. The dataset covers diverse topics across a range of subjects. Moreover, we do not focus only on major school subjects such as Chemistry, Physics, Mathematics, and Biology, but also cover highly specialized ones such as Agriculture, Geology, Informatics, and History. The goal of this dataset is to evaluate and reliable development of multimodal models that are capable of doing multimodal parsing of complex textual and visual information present in an image and arrive at the correct answer.

3.1 Data Collection and Analysis

In this section, we present the properties of the dataset, and we give details about the process of data collection, preparation, and normalization, as well as information about the data splits.

Collection and Preparation of Dataset We collect Exams from official state and national exams prepared by the ministries and national education agencies of various countries. These exams are taken by students graduating from high school and often require knowledge learned through the entire course. We identified potential online sources of publicly available school exams and previous year's paper sources of National exam agencies. We downloaded the PDF files of the exam paper from these sources per year, per subject.

Then, we wanted to get snapshots of questions from these PDFs. To do so, we converted the PDFs into high-quality images by pages. Then we used an open-sourced labeling pipeline to manually annotate bounding boxes for cropping out questions from each image. While annotating, we restricted ourselves to annotating only multi-choice questions with one single option. To make the evaluation and dataset simpler, we avoid including questions that have free-form answers, multi-choice questions with more than one correct answer, Interger-type questions, and Numeric-type questions.

Data Statistics The main statistics of Exams are presented in Table 1. There are 14383 samples in total of which 2341 samples have only textual context and 9407 samples have both visual and textual context. From the collected data we have carefully curated a test set and validation set. The test set consists of two partitions: the first partition has questions with both visual and text context and the second partition has questions with only text context. In each partition, for every language, there are 50 questions for every subject present in the dataset. This ensures that the evaluation test dataset is unbiased towards a particular language-subject combination. The remaining data can be used to create a training and validation set or to create an in-context learning sampling set during inference.

Language Diversity Our dataset includes a total of 14383 samples in 11 languages from 6 different language families. Each question is a 3-way to

language	Family	grade	no of questions	no of questions with image	no of text only
Bulgarian	Balto-Slavic	4	497	42	455
		12	1,635	460	1175
Chinese	Sino-Tibetan	12	2,635	0	0
Croatian	Balto-Slavic	12	3,973	744	3229
French	Romance	12	439	66	373
German	Germanic	12	819	174	645
Hungarian	Finno-Ugric	12	3,801	695	3106
Italian	Romance	12	44	4	40
Romanian	Romance	12	5	0	5
Russian	East-Slavic	12	9	0	9
Serbian	Balto-Slavic	12	227	67	160
Spanish	Romance	12	299	89	210

Table 1: Statistics table

5-way multiple-choice question with a single correct answer. Table 1 shows a breakdown of each language, where **no of subjects**, no of samples, and their breakdown to text only and visual and text samples are shown in detail. This dataset has two language family groups from the Slavic family and the Romance family 4 language representations from each family. There is also a distant language family like the Sino-Tibetan language family with Chinese as its representative. Finally, we have representatives from high and low-resource languages in our dataset. All these characteristics make our dataset an ideal fit for a multimodal and multilingual assessment of any multimodal model or pipeline with visual question-answering capabilities.

Domain Diversity Each education system has its own specifics, resulting in some differences in curricula, topics, and even naming of the subjects. Given the sparse nature of the subjects, we use a two-level taxonomy in order to categize them into logically connected groups. The lower-level is a subject and the higher level is a major group. We normalized the subject using a two-step algorithm: first, we put each subject in a separate category, then, if the subject was general enough, e.g. Biology, History, etc., or there were no similar ones, we retained the category: otherwise, we merged all similar subjects together in a unifying category, e.g., Economics Basics, and Economics & Marketing. We repeated the aforementioned steps until there were no suitable merge candidates. As a result, we ended up with a total of — subjects. We further grouped into three major categories, based on the main branches of science: Natural Science - *"the study of natural phenomena"*, Social Science = *"the study of human behavior and societies"*, Other - *Applied Studies, Arts, Religion etc.*.

Question Complexity The dataset is created from high-school examinations conducted in different countries. Most of the questions are from grade 12 while Bulgarian has few questions from grade 4. To answer these questions correctly, the model requires both complex reasoning, and a fundamental understanding of physics, chemistry, biology, and mathematics. Also, there a questions from geography and history that require specific information of a particular region or country.

3.2 Comparision with Existing Dataset

4 Experiments

4.1 Language Model-based Pipeline

5 Experiments and Results

6 Analysis

7 Discussion

References

A Example Appendix

This is an appendix.