DATA INTERPRETATION AND REASONING OVER SCIENTIFIC PLOTS

Anonymous authors
Paper under double-blind review

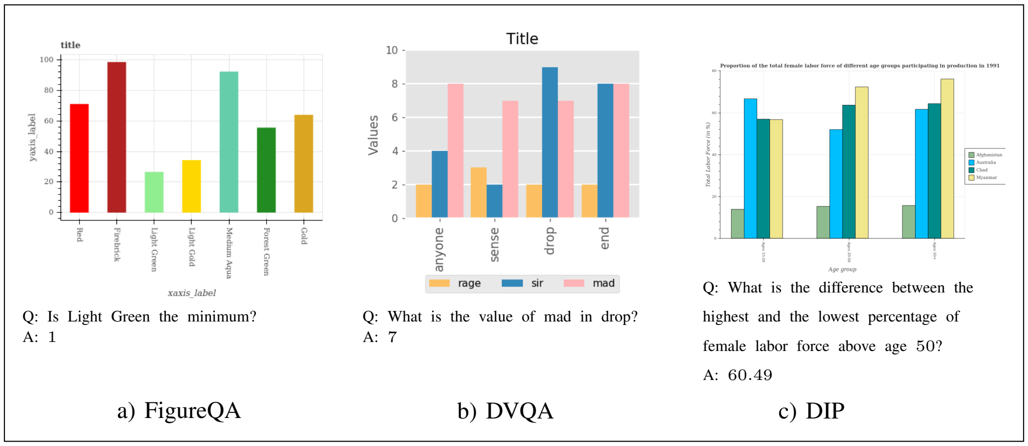
ABSTRACT

Data Interpretation is an important part of Quantitative Aptitude exams and requires an individual to answer questions grounded in plots such as bar charts, line graphs, scatter plots, *etc*. Recently, there has been an increasing interest in building models which can perform this task by learning from datasets containing triplets of the form {plot, question, answer}. Two such datasets have been proposed in the recent past which contain plots generated from synthetic data with limited (i) $x - y$ axes variables (ii) question templates and (iii) answer vocabulary and hence do not adequately capture the challenges posed by this task. To overcome these limitations of existing datasets, we introduce a new dataset containing 9.7 million question-answer pairs grounded over $270,000$ plots with three main differentiators. First, the plots in our dataset contain a wide variety of realistic x-y variables such as CO2 emission, fertility rate, *etc*. extracted from real word data sources such as World Bank, government sites, *etc*. Second, the questions in our dataset are more complex as they are based on templates extracted from interesting questions asked by a crowd of workers using a fraction of these plots. Lastly, the answers in our dataset are not restricted to a small vocabulary and a large fraction of the answers seen at test time are not present in the training vocabulary. As a result, existing models for Visual Question Answering which largely use end-to-end models in a multi-class classification framework cannot be used for this task. We establish initial results on this dataset and emphasize the complexity of the task using a multi-staged modular pipeline with various sub-components to (i) extract relevant data from the plot and convert it to a semi-structured table (ii) combine the question with this table and use compositional semantic parsing to arrive at a logical form from which the answer can be derived. We believe that such a modular framework is the best way to go forward as it would enable the research community to independently make progress on all the sub-tasks involved in plot question answering.

1 INTRODUCTION

Data plots such as bar charts, line graphs, scatter plots, etc. provide an efficient way of summarizing numeric information and are frequently encountered in textbooks, research papers, professional reports, newspaper articles, *etc*. Machine comprehension of these plots with the aim of answering questions grounded in them is an interesting research problem which lies at the intersection of Language and Vision and has widespread applications. For example, such a system could enable financial analysts to use natural language questions to access the information locked in a collection of data plots within financial reports, journals, etc. Such a system could also serve as an important educational assistant for the visually impaired by helping them understand the information summarized in a plot by asking a series of questions.

Recently, two datasets , *viz*., FigureQA(Kahou et al., 2017) and DVQA(Kafle et al., 2018) have been released for this task which contain triplets of the form {plots, questions, answers}. These datasets clearly show that despite its apparent similarity to Visual Question Answering (VQA), this task has several additional challenges due to which existing state of the art VQA methods do not perform well on this task. However, both FigureQA and DVQA themselves have some limitations which warrants the creation of more challenging datasets which adequately capture a wider range of



Q: Is Light Green the minimum?
A: 1

a) FigureQA

Q: What is the value of mad in drop?
A: 7

b) DVQA

Q: What is the difference between the highest and the lowest percentage of female labor force above age 50?
A: 60.49

c) DIP

Figure 1: A sample {plot, question, answer} triplet from FigureQA, DVQA and DIP datasets.

challenges posed by this task. With this motivation, we create a new dataset for Data Interpretation over Plots (DIP) with three major differentiators from existing datasets as outlined below.

Figure 1 shows a sample triplet from FigureQA, DVQA and DIP (our dataset). First, we note that FigureQA and DVQA contain plots created from synthetic data. In particular, note that the label names (either x-axis or y-axis or legend names) in FigureQA and DVQA come from a limited vocabulary (*color names* and *top-1000 nouns* from Brown corpus respectively). This clearly reduces the vocabulary that a model needs to deal with. Further, the label names are not really meaningful in the context of the plot leading to unnatural questions. In contrast, the plots in our dataset are based on World Bank Open Data which contains realistic variable names such as mortality rate, crop yield, country names, *etc*. The values associated with each plot are also realistic with different scales including floating point numbers as opposed to DVQA which contains only integers in a fixed range. Secondly, the questions in FigureQA and DVQA are based on a smaller number of templates (15 and 25). Instead of hand designing a small number of templates, we first show a fraction of the plots to crowdsourced workers and ask them to create natural language questions which can be answered from these plots. We then analyze these questions and extract templates from them which results in a richer question repository with more templates (74) and more complex questions. Lastly, unlike FigureQA and DVQA the answers in our dataset do not come from a small fixed set of vocabulary. For example, the answer to the question shown in Figure 1 is 60.49, which is not present in the training data. More specifically, the answer vocabulary for the test data is $248,878$ words of which $187,688$ are not seen in the training data. This is quite natural and expected when the plots and questions are extracted from real world data. In addition to the above differentiators, we also include an extra novel test set which contains plots based on data extracted from Open Government Data as opposed to World Bank Data. This test set contains additional variable and legend names which are not seen in the World Bank Data.

Given the large answer vocabulary, it is infeasible to use any of the existing VQA models on our dataset as they largely treat VQA as a multi-class classification problem where the task is to select the right answer from a fixed vocabulary. Even the recent models proposed on DVQA take a similar multi-class classification approach. Further, we believe that given the various sub-tasks involved in this problem it is not prudent to use a single end-to-end system which simply takes the plot and question as input and generates as answer. Instead, as a baseline we propose a modular multi-staged pipeline wherein the first stage in the pipeline extracts (i) data objects in the plot such as bars, lines, *etc*. (ii) text objects in the plot such as titles, legend names, x-y axes names, *etc.* (iii) the numeric objects in the plot such as tick values, scales, etc. At the end of this stage, the plot is converted to a semi-structured table. We then use a method (Pasupat & Liang, 2015) which combines the question with this table and uses compositional semantic parsing to arrive at a logical form from which the answer can be derived. The key point here is that the output is neither selected from a fixed vocabulary nor generated using a decoder but it is derived from a logical form. Our experiments using this model suggest that this dataset is indeed very challenging and requires the community to look beyond existing end-to-end models.

2 Related Work

In this section, we review existing datasets and models for Visual QA.

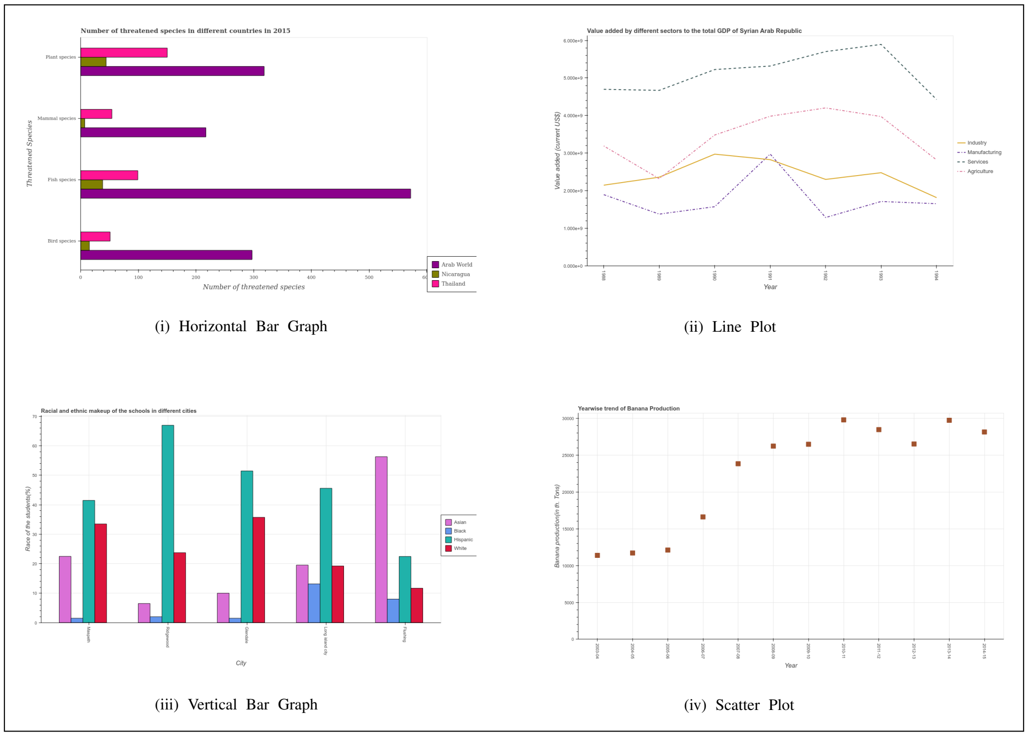
Datasets: Over the past few years several large scale datasets for Visual Question Answering have been released. These include datasets such as COCO-QA (Ren et al., 2015), DAQUAR (Malinowski & Fritz, 2014), VQA (Antol et al., 2015; Goyal et al., 2017) which contain questions asked over natural images. On the other hand, datasets such as CLEVR (Johnson et al., 2017) and NVLR (Suhr et al., 2017) contain complex reasoning based questions on synthetic images having 2D and 3D geometric objects. There are some datasets (Kembhavi et al., 2016; 2017) which contain questions asked over diagrams found in text books but these datasets are smaller and contain multiple-choice questions. Another dataset worth mentioning is the FigureSeer dataset (Siegel et al., 2016) which contains images extracted from research papers but this is also a relatively smaller dataset (60,000 images). Further, their focus is on extracting information and answering questions based on line plots as opposed to other types of plots such as bar charts, scatter plots, *etc.* as seen in two recently released datasets, *viz.*, FigureQA (Kahou et al., 2017) and DVQA (Kafle et al., 2018). However, these two datasets also have some limitations which we try to overcome by proposing a new dataset in this work. To the best of our knowledge, ours is the first work which introduces plots created from real-world data containing natural x-y axes variables, real data values, a huge answer vocabulary and richer question templates.

Models: The availability of the above mentioned datasets has facilitated the development of complex end-to-end neural network based models (Lu et al. (2016), Yang et al. (2016), Noh & Han (2016), Santoro et al.), which typically contain (i) encoders to compute a representation for the image and the question (ii) attention mechanisms to focus on important parts of the question and image (iii) interaction components to capture the interactions between the question and the image and finally (iv) a classification layer for selecting the answer from a fixed vocabulary. While these algorithms have shown reasonable success on the original VQA tasks that they were proposed for, they fail to perform well on datasets where the questions involve more complex numeric reasoning over images. A case in point is the performance of the models on the recently released datasets for plot QA, *viz.*, FigureQA (Kahou et al., 2017) and DVQA (Kafle et al., 2018). One crucial difference is that when dealing with plots such as bar charts, scatter plots, etc. the numeric information in the image needs to be accurately inferred based on the position of the axes, relative position of legends, bars, lines etc. in the figure. Such structure is typically not present in natural images and the resulting questions are also more complex and quantitative in nature. An additional challenge in our dataset is that the answer vocabulary is not fixed and can range from floating point numerical answers to textual answers containing multiple words. As a result, existing end-to-end models which treat VQA as a multi-class classification problem cannot be used. We propose that instead of using a single end-to-end model which addresses all the sub-tasks of plot QA such as extracting label information, numeric information, parsing questions, *etc.* one should build a modular pipeline wherein different modules address different sub-tasks involved. In particular, we first convert the plot to a semi-structured table (similar to (Cliche et al., 2017)) and then leverage existing methods (Berant et al.) for answering questions from such tables (Pasupat & Liang, 2015) using compositional semantic parsing.

3 The DIP (Data Interpretation over Plots) Dataset

In this section, we describe the four main stages involved in building this dataset, *viz.*, (i) curating data such as year-wise rainfall statistics, country-wise mortality rates, *etc.* (ii) creating different types of plots with a variation in the number of elements, legend positions, fonts, etc. (iii) using crowdsourcing to generate questions (iv) extracting templates from the crowdsourced questions and instantiating these templates using appropriate phrasing suggested by human annotators.

We first describe in detail the various phases that were involved in creating the dataset.



(i) Horizontal Bar Graph

(ii) Line Plot

(iii) Vertical Bar Graph

(iv) Scatter Plot

Figure 2: Sample of each of the plot types present in the DIP dataset.

3.1 Data Collection and Curation

We considered online data sources such as World Bank Open Data[1], Open Government Data[2], Global Terrorism Database[3], Nutritional Analysis Data[4], *etc.* which contain statistics about various indicator variables such as fertility rate, rainfall, coal production, *etc.* across years, countries, districts, *etc.* We crawled data from these sources to extract different variables whose relations could then be plotted (for example, rainfall v/s years or rainfall v/s years across countries or movie v/s budget or carbohydrates v/s food_item and so on). Some important statistics about the crawled data are as follows: (i) it has 841 unique indicator variables (ii) the data ranges from 1960 to 2016 (of course, all indicator variables may not be available for all years) and (iii) the data corresponds to 160 unique entities (cities, states, districts, countries, movies, food items, etc.). The data contains positive integers, floating point values, percentages and values on a linear scale. The minimum value across all indicator variables is 0 and the maximum value is $3.50E + 15$.

3.2 Plot Generation

We included 3 different types of plots in this dataset, *viz.*, bar graphs, line plots and scatter plots. Within bar graphs, we have grouped bar graphs with both horizontal and vertical orientation. These types cover a wide range of plots that are typically encountered in journals, research papers, textbooks, *etc.* We couldn't find enough data to create certain other types of plots such as Venn diagrams and pie charts which are used in very unique situations. We also do not consider composite plots such as Pareto charts which have line graphs on top of bar graphs. Lastly, all the plots in our dataset contain only 2-axes. Figure 2 shows one sample from each of the plot types in our dataset. Note that in Figure 2 (iii), the plot compares an indicator variable (say, race of students) across cities for different racial origins. Each plot is thus a compact way of representing 3 dimensional data. To enable

[1] https://data.worldbank.org/

[2] https://www.india.gov.in/

[3] https://www.start.umd.edu/gtd/

[4] https://www.kaggle.com/mcdonalds/nutrition-facts/home

the development of supervised modules for various sub-tasks we provide bounding box annotations for legend boxes, legend names, legend markers, axes titles, axes ticks, bars, lines and plot title. By using different combination of indicator variables and entities (years, countries, *etc.*) we created a total of $273,821$ plots.

To ensure that there is enough variety in the plots, the following plot parameters were randomly chosen: (i) grid lines (present/absent) (ii) font size (iii) the notation used for tick labels (scientific-E notation or standard notation (iv) line style in the case of line plots, *i.e*, $\{$solid, dashed, dotted, dash-dot$\}$ (v) marker styles for marking data points, *i.e.*, $\{$asterisk, circle, diamond, square, triangle, inverted triangle$\}$ (vi) position of legends, *i.e.*, $\{$bottom-left, bottom-centre, bottom-right, center-right, top-right$\}$ and (vii) colors for the lines and bars from a set of 73 unique colors. The number of discrete elements on the x-axis varies from 2-12. Similarly, the number of entries in the legend box varies from 1-4. In other words, in the the case of line plots, the number of lines varies from 1-4 and in the case of grouped bars the number of bars grouped on a single x-tick varies from 1-4. For example, for the plots in Figure 2 (iii), the number of discrete elements on the x-axis is 5 and the number of legend names (*i.e.*, number of lines or number of bars in a group) is 4.

3.3 Sample Question Collection via Crowdsourcing

Since the underlying data in our plots is much richer as compared to FigureQA and DVQA we felt that instead of creating a small number of templates it would be best to ask a wider set of annotators to create questions over these plots. However, creating questions for all the plots in our dataset would have been prohibitively expensive. To avoid this, we sampled $3,500$ plots across different types and asked workers on Amazon Mechanical Turk to create questions for these plots. We showed each plot to 5 different workers resulting in a total of $17,500$ questions. We specifically instructed the workers to ask complex reasoning questions which involved reference to multiple plot elements in the plots. We also gave the workers a list of simple questions such as “Is this a vertical bar graph?”, “What is the title of the graph?”, ”What is the value of coal production in 1993?” and asked them to strictly not create such questions as we had already created such questions using hand designed templates. We paid them 0.1\$ for each question.

3.4 Question Template Extraction and Instantiation

We manually analyzed the questions collected via crowdsourcing and divided them into 5 coarse classes with a total of 74 templates (including the simple templates that we had manually designed as mentioned earlier). These question categories along with a few sample templates are shown below (we refer the reader to the Supplementary material for more details).

1. **Structural Understanding** : These are questions about the overall structure of the plot and typically do not require any quantitative reasoning. For example, “How many different coloured bars are there ?”, “How are the legend labels stacked ?” and so on.
2. **Data Retrieval** : These questions are typically related to a single element in the plot. For example, “What is the value added by services sector in 1992 ?”, “How many bars are there on the 4^{th} tick from the top ?” and so on.
3. **Numeric Reasoning** : These questions require some numeric reasoning over multiple plot elements. For example, “Across all countries, what is the maximum number of threatened fish species ?”, “In which country was the number of threatened bird species minimum ?”, “What is the median banana production ?”, “What is the difference between the number of threatened bird species in Thailand and that in Nicaragua ?” and so on.
4. **Comparative Reasoning** : These questions typically require a comparative analysis of different elements of the plot. For example, “In how many countries, is the number of threatened fish species greater than 180 ?”, “What is the ratio of the value added by industrial sector in 1988 to that in 1989 ?” and so on.
5. **Compound** : These questions typically combine numeric and comparative reasoning. For example, “What is the difference between the highest and the second highest banana production ?”, “In how many years, is the banana production greater than the average banana production taken over all years ?” and so on.

We also created corresponding logical forms for these questions which could be executed on the raw data to extract the answers for the questions. Generating questions for multiple indicator variables using these templates was a tedious task requiring a lot of manual intervention. For example, consider the template "In how many <plural form of X_label>, is the <Y_label> of/in <legend_label> greater than the average <Y_label> of/in <legend_label> taken over all <plural form of X_label> ?". Now consider the indicator variable "Race of students" in Figure 2 (iii). If we substitute this indicator variable as it is in the question it would result in a question, "In how many cities, is the race of the students(%) of Asian greater than the average race of the students (%) of Asian taken over all cities ?", which sounds unnatural. To avoid this, we had to carefully paraphrase these indicator variables and question templates so that the variables could be automatically substituted in the template. As a result, when each template is instantiated for different indicator variables it looks a bit different in its surface form.

4 Proposed Pipeline

We break down the task of Plot QA into two main tasks. The first task is to extract all numeric/textual data from the plot and store it in a semi-structured table. Once we have such a table, the next subtask is to perform QA over such a table as opposed to the image directly. The task of converting the plot to a semi-structured table can itself be broken down into multiple sub-tasks. The modules used to perform all these sub-tasks as well as the final Table-QA are shown in Figure 3 and described below.

4.1 Visual Elements Detection (VED)

The information contained in plots is very sparse as most of the pixels simply belong to the background and very few pixels/locations contain actual data. More specifically, the data bearing elements of the plot are the title of the plot, the labels of the x and y axes, the tick marks or categories (e.g., countries) on the x and y axis, the data markers in the legend box, the legend names, the full legend box and finally the actual bars and lines in the graph. Following existing literature (Cliche et al. (2017),Kafle et al. (2018)), we refer to these elements as the visual elements of the graph. The first task is to extract all these visual elements by drawing bounding boxes around them. We treat this as an object detection task and use the Single Shot Multibox Detector (SSD) (Liu et al., 2016) to detect these elements in the graph. We train one such object detector for each of the 9 visual elements and run them in parallel at inference time to detect all the elements in the plot.

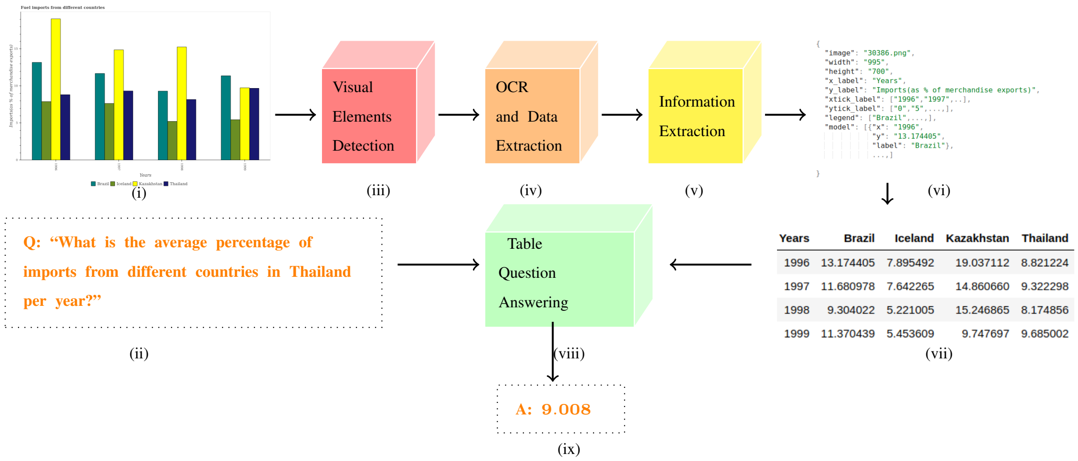
4.2 Optical Character Recognition (OCR) and Data Extraction

Once we have the bounding boxes, the next task is to read the numeric or textual values inside these bounding boxes. This is a task of Optical Character Recognition(OCR) and we use a state of the art OCR model (Smith, 2007) for this task. Specifically, given the coordinates of the bounding box, we crop the image to this bounding box, convert the cropped image into grayscale, resize and deskew it and then pass it to an OCR module. Since existing OCR modules perform very well for machine written English text, we simply use an existing pre-trained OCR module [5] and do not re-train it using our data.

4.3 Semi-Structured Information Extraction (SIE)

As shown in Figure 3, using all the information extracted so far, we are interested in creating a table where the rows correspond to the ticks on the x-axis (1996, 1997, 1998, 1999 in this case), the columns correspond to the different elements listed in the legend (Brazil, Iceland, Kazakhistan, Thailand in this case) and the i,j-th cell contains the value corresponding to the x-th tick and the y-th legend. The values of the x-tick labels and the legend names are already available from the OCR module. However, we still need to identify which legend name corresponds to which legend marker (or color in this case). More specifically, in the above example, we have already identified the bounding boxes of all the legend markers (colors in this case) and the legend names but we need to now associate the right color to the right legend name. We do this by simply associating

[5] https://github.com/tesseract-ocr/tesseract


Fuel imports from different countries
Imports(as % of merchandise exports)
0
5
10
15
1996
1997
1998
1999
Years
Brazil
Iceland
Kazakhstan
Thailand
(i)
Visual Elements Detection
(iii)
OCR and Data Extraction
(iv)
Information Extraction
(v)
(vi)
Q: “What is the average percentage of imports from different countries in Thailand per year?”
(ii)
Table Question Answering
(viii)
(vii)
A: 9.008
(ix)

```
{
  "image": "30386.png",
  "width": "995",
  "height": "700",
  "x_label": "Years",
  "y_label": "Imports(as % of merchandise exports)",
  "xtick_label": ["1996","1997",..],
  "ytick_label": ["0","5",...,],
  "legend": ["Brazil",...,],
  "model": [{"x": "1996",
             "y": "13.174405",
             "label": "Brazil"},
             ...,]

}
```

Years	Brazil	Iceland	Kazakhstan	Thailand
1996	13.174405	7.895492	19.037112	8.821224
1997	11.680978	7.642265	14.860660	9.322298
1998	9.304022	5.221005	15.246865	8.174856
1999	11.370439	5.453609	9.747697	9.685002

Figure 3: Our proposed multi-staged modular pipeline which (i) takes the image of the plot and (ii) question as input, (iii) extracts bounding boxes for all visual elements of the plot, (iv) applies OCR to extract the textual and numerical content of the visual elements, (v) associates the data-points with appropriate legend labels, x and y co-ordinates, (vi) consolidates all the information about the input plot into JSON format, (vii) converts the json formatted file into a semi-structured table. (viii) Further, the Question Answering module which takes the question and the semi-structured table and predicts the (ix) answer.

a legend name to the marker whose bounding box is closest to the bounding box of the legend name. Similarly, we associate each tick label to the tick marker whose bounding box is closest to the bounding box of the tick label. In other words, we have now associated, the legend name Brazil to the color “Dark Cyan” and the tick label 1996 to the corresponding tick mark on the x- axis.

Once we have identified the 4 row headers (years, in this case) and the 4 column headers (countries, in this case) we need to fill a total of 16 values in the table. This is a two step process, wherein we first need to associate each of the 16 bounding boxes corresponding to the 16 bars to the corresponding x-tick and legend name. Associating a bar to the corresponding x-tick label is again straightforward as we can simply associate it with the tick whose bounding box is closest to the bounding box of the bar. To associate a bar to a legend name we find the dominant color in the bounding box of the bar and associate it with the legend name associated with that color. Lastly, we need to find the value represented by each bar. For this, we use the tick marks on the y-axis as identified by the OCR module as a reference. In particular, we extract the height of the bar using the bounding box information and then look at the y-tick labels immediately above and below it. We can then interpolate the value of the bar based on the values of these ticks.

At the end of this stage we have extracted all the information from the plot into a semi-structured table as shown in Figure 3 (vii).

4.4 TABLE QUESTION ANSWERING (QA)

Once we have the data in a semi-structured table the task becomes similar to the task of answering questions from the WikiTableQuestions dataset (Pasupat & Liang, 2015). We simply use the approach outlined in Pasupat & Liang (2015) wherein they first convert the table to a knowledge graph, convert the question to a set of candidate logical forms by applying compositional semantic parsing. Each of these logical forms is then ranked using a log-linear model and the highest ranking logical form is then applied to the knowledge graph to get the answer. Due to space constraints, we refer the reader to the original paper for more details of this stage.

Table 1: DIP Dataset Statistics

Dataset Split	**#Images**	**#QA pairs**
Train	1,57,070	5,733,893
Validation	33,650	1,228,468
Test-Familiar	33,657	1,228,313
Test-Novel	48,444	1,593,138
Test-Hard	1,000	3,000
Total	273,821	9,786,812

Table 2: Accuracy of the Semi-Structured Information Extraction(SIE) module on the various dataset-splits with varying values of threshold.

Threshold	**Test-Familiar**	**Test Novel**
1%	7.9%	5.6%
10%	58.53%	53.23%

Table 3: Arithmetic Mean accuracy calculated by computing the Levenshtein distance between the ground-truth label and the label predicted after doing OCR.

Dataset Split	**Test-Familiar**	**Test-Novel**
Title	96.3%	97.1%
X-Label	97.8%	99.6%
Y-Label	85.2%	85.7%
Legend Label	87.7%	92.1%
Tick Label	93.4%	90.1%
Total	92.9%	90.7%

Table 4: Average Precision (AP) of the SSD models for different classes present in the plots of the DIP dataset

Class	Test-Familiar	Test-Novel
Title	99.4%	98.9%
Bar	62.3%	63.5%
Line	45.8%	60.1%
Scatter	82.6%	89.7%
X-axis Label	99.4%	99.3%
Y-axis Label	28.7%	10.3%
X-tick Label	98.4%	98.8%
Y-tick Label	96.4%	96.5%
X-tick Mark	16.5%	15.9%
Y-tick Mark	5.3%	3.9%
Legend Label	88.1%	92.8%
Legend Preview	84.3%	91.1%
mAP	**67.27%**	**68.41%**

5 Experimental Setup

Train-Valid-Test Splits: As mentioned earlier, by using different combinations of 841 indicator variables and 160 entities(years, countries, *etc*), we created a total of 273, 821 plots. Depending on the context and type of the plot, we instantiated the 74 templates to create meaningful (questions,answer) pairs for each of the plots. The number of questions per plot varies from 17-44. We created train (58%), valid(12%), test-familiar (12%) and test-novel (17%) splits from this data. The difference between test-familiar and test-novel is that test-novel contains plots from a data source (Open Government Data) which was not seen at training time and hence has almost no overlap with the indicator variables seen at training time. We also created a test-hard split which contains 3000 questions created by crowdsourced workers from 1000 plots. However, we do not use the test-hard split for our experiments. These statistics are summarized in Table 1.

Training Details: Of the 4 components described in Section 4, only two require training, *viz.*, Visual Elements Detection (VED) and Table Question Answering(QA). As mentioned earlier, for VED we need to train one object detection model (SSD (Liu et al., 2016)) for each of the 9 visual elements. We observed that if we use a single object detection model for all the elements then the performance drops significantly. These models were trained using the bounding box annotations available in our dataset. We trained each model for 200, 000 steps with a batch size of 32. We did not see any benefit of training beyond 200, 000 steps. We used RMSrop[6] as the optimizer with an initial learning rate of 0.004. Next, for the Table QA module we trained the model proposed in (Pasupat & Liang, 2015) using questions from our dataset and the corresponding ground truth tables which were known to us. Since this model is computationally expensive with a high training time, we could train it using only 200, 000 questions from our training set.

[6] http://www.cs.toronto.edu/ tijmen/csc321/slides/lecture_slides_lec6.pdf

Table 5: Accuracy of Compositional Semantic Parsing on DIP dataset for each question category.

Dataset Split	**Validation (Ground-truth)**	**Validation (Predicted)**	**Test-Familiar (Predicted)**	**Test-Novel (Predicted)**
Structural	13.17%	13.21%	13.37%	5.97%
Data Retrieval	54.8%	26.8%	26.83%	27.86%
Numeric Reasoning	67.20%	39.02%	39.48%	17.4%
Comparative Reasoning	23.7%	18.19%	18.53%	6.15%
Compound	1.29%	1.11%	1.2%	1.9%
Total	**32.9%**	**19.9%**	**20.2%**	**13.10%**

6 Results

We now discuss the performance of each of the components of our pipeline.

- **VED Module**: We evaluate this module by computing the average precision of the detected bounding box for all the visual elements. We use the same methodology as outlined in the original PASCAL-VOC 2012 challenge[7]. We report the per element average precision (AP) in Table 4. We observe that the object detection module does well for most elements except line, y-label, x-tick and y-tick. On further investigation, we found that x-tick and y-tick are typically very small and thus hard to detect. Similarly, line elements are very thin and hence hard to detect. Lastly, we found that the Y-axis label is often very close to the tick values which makes it hard to detect. While overall the performance of this module is satisfactory, significant improvements are needed to use it in an end-to-end plot QA pipeline.
- **OCR Module** : For all the textual elements, we took the bounding boxes predicted by VED and then passed them through a pre-trained OCR engine. We then computed the accuracy of OCR as the percentage of characters which were detected correctly. We report this accuracy in Table 3 and observe that the OCR module performs reasonably given that it is also affected by the errors of the VED module. This suggests that going forward, we could pay less attention to this module and focus on the other components in the pipeline.
- **SIE Module** : As mentioned earlier, we have the ground truth table corresponding to each plot. To compute the performance of this module we compare each entry in the table extracted by this module with the corresponding entry in the ground-truth table. We consider the two entries to match if the predicted value is within k% of the true value. We refer to k as the threshold. Table 2 summarizes the performance of the SIE module for different threshold values. This module clearly needs further improvement. Currently, this is a heuristic based module and there is clear scope for improving it using a learning based algorithm.
- **QA Module** : The task of this module is to find the answer based on the table extracted by SIE and the given question. We first evaluate the standalone accuracy of this module on the validation set when the table is the ground-truth table. We observe that even when the table is a ground truth table *i.e.* even when all the entries in the table and row and column names are correct, this module only achieves an accuracy of 32.9% (see Table 5). As expected, when we replace the ground truth with that predicted by SIE after taking the output from OCR and VED, the accuracy drops significantly. The last 3 columns of this table thus report the end-to-end accuracy on different splits. The results are clearly low and we clearly need to fcous on this last stage of the pipeline. In particular, we observe that the results are very bad for questions requiring comparative and compound reasoning.

7 Conclusion

We introduced the DIP dataset for Data Interpretation over Plots which contains scientific plots created from real world data sources such as World Bank, stock market, *etc.* Further, the questions in our dataset are based on templates which were manually extracted from realistic questions created

[7] http://host.robots.ox.ac.uk/pascal/VOC/voc2010/devkit_doc_08-May-2010.pdf

by crowd-workers. One of the primary challenges of this dataset is that it has a large vocabulary because of which existing VQA models which treat this as a multi-class classification model cannot be applied. Instead we propose a modular pipeline which first converts the plot to a semi-structured table and then learns to answer questions from this table using compositional semantic parsing. Our experiments suggest that this is a very challenging dataset and requires significant progress on multiple sub-tasks. In particular, we need improved models for reasoning over structured data.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Mathieu Cliche, David S. Rosenberg, Dhruv Madeka, and Connie Yee. Scatteract: Automated extraction of data from scatter plots. In *ECML PKDD*, 2017.

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 *CVPR*, 2017.

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. In *CVPR*, 2017.

Kushal Kafle, Scott Cohen, Brian L. Price, and Christopher Kanan. DVQA: understanding data visualizations via question answering. *CoRR*, abs/1801.08163, 2018.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *CoRR*, abs/1710.07300, 2017.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Aniruddha Kembhavi, Min Joon 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 *CVPR*, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.

Hyeonwoo Noh and Bohyung Han. Training recurrent answering units with joint loss minimization for VQA. *CoRR*, abs/1606.03647, 2016.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *ACL*, 2015.

Mengye Ren, Ryan Kiros, and Richard S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR*, abs/1505.02074, 2015.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *NIPS*.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Kumar Divvala, and Ali Farhadi. Figureseer: Parsing result-figures in research papers. In *ECCV*, 2016.

Ray Smith. An overview of the tesseract ocr engine. In *ICDAR*, 2007.

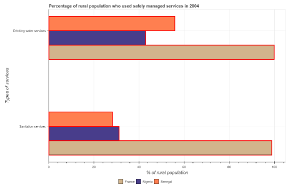
Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

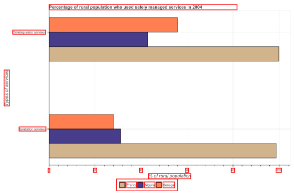
APPENDICES

A. EXAMPLE PLOTS

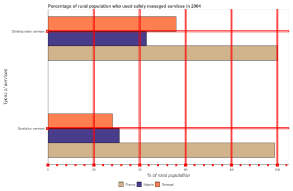
Few instances of the {plot, question, answer} triplet are given below. We have also shown the bounding box annotations which is used to recognise every visual element in the plot.



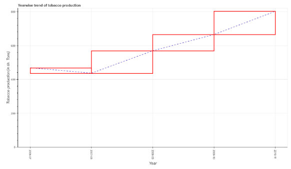
Q: What is the percentage of rural population who used sanitation services in Senegal ?
A: 28.2



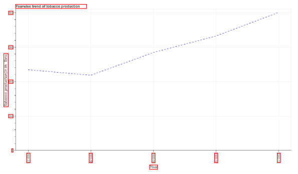
Q: In how many countries, is the percentage of rural population who used drinking water services greater than the average percentage of rural population who used drinking water services taken over all countries ?
A: 1



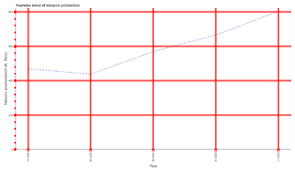
Q: Are the values on the major ticks of X-axis written in scientific E-notation ?
A: No



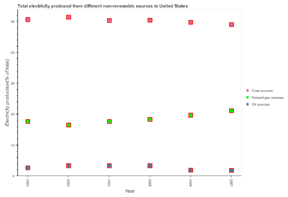
Q: What is the average tobacco production per year ?
A: 588



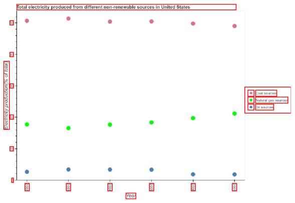
Q: What is the ratio of the tobacco production in 2007-08 to that in 2009-10 ?
A: 0.657



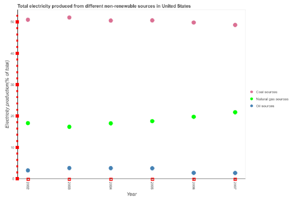
Q: Does the tobacco production monotonically increase over the years ?
A: No



Q: What is the difference between two consecutive major ticks on the Y-axis ?
A: 10



Q: Is the sum of the percentage of electricity produced by oil sources in 2005 and 2007 greater than the maximum percentage of electricity produced by coal across all years ?
A: No



Q: In how many years, is the percentage of electricity produced by natural gas greater than 50%?
A: 0

B. PLOT STATISTICS

Table 6 presents the detailed statistics of the number of plots present in each of our data splits according to plot type. Note that the plot type k-multi means that the number of lines/bars on each tick is k. In the above mentioned samples, the horizontal bar graph is 3-multi and so on.

Table 6: Detailed statistics of the dataset for different splits for different figure types

Dataset Split	Plot Type	vbar	hbar	line	scatter
Train	single	8,906	8,825	5,285	5,315
	2-multi	9,434	9,756	5,047	5,091
	3-multi	18,772	18,756	8,295	8,309
	4-multi	15,351	15,363	7,270	7,295
Validation	single	1,908	1,891	1,132	1,138
	2-multi	2,021	2,090	1,081	1,090
	3-multi	4,022	4,019	1,777	1,780
	4-multi	3,298	3,292	1,557	1,563
Test-Familiar	single	1,908	1,891	1,132	1,139
	2-multi	2,021	2,090	1,081	1,091
	3-multi	4,023	4,019	1,778	1,781
	4-multi	3,290	3,292	1,558	1,563
Test-Novel	single	10,604	10,610	3,446	3,565
	2-multi	4,862	4,874	1,243	1,240
	3-multi	1,850	1,850	900	900
	4-multi	1,000	1,000	250	250

C. Question Templates

The templates which we have used for question generation are given below. Note that, not all question templates are applicable to each and every type of plot. Also depending on the context of the plot, the question varies largely along the template's surface form.

1. **Structural Understanding** :
   (a) Does the graph contain any zero values?
   (b) Does the graph contain grids ?
   (c) Where does the legend appear in the graph ?
   (d) How many legend labels are there?
   (e) How are the legend labels stacked?
   (f) How many <plural form of X_label> are there in the graph?
   (g) How many <figure-type>s are there?
   (h) How many different colored <figure-type>s are there?
   (i) How many groups of <figure-type>s are there?
   (j) Are the number of bars on each tick equal to the number of legend labels?
   (k) Are the number of bars in each group equal?
   (l) How many bars are there on the i^{th} tick from the left?
   (m) How many bars are there on the i^{th} tick from the right?
   (n) How many bars are there on the i^{th} tick from the top?
   (o) How many bars are there on the i^{th} tick from the bottom?
   (p) Are all the bars in the graph horizontal?
   (q) How many lines intersect with each other?
   (r) Is the number of lines equal to the number of legend labels?
2. **Data Retrieval** :
   (a) What does the i^{th} bar from the left in each group represent?
   (b) What does the i^{th} bar from the right in each group represent?
   (c) What does the i^{th} bar from the top in each group represent?
   (d) What does the i^{th} bar from the bottom in each group represent?
   (e) What is the label of the j^{th} group of bars from the left?
   (f) What is the label of the j^{th} group of bars from the top?

(g) Does the <Y_label> of/in <legend-label> monotonically increase over the <plural form of X_label> ?

(h) What is the difference between two consecutive major ticks on the Y-axis ?

(i) Are the values on the major ticks of Y-axis written in scientific E-notation ?

(j) What is the title of the graph ?

(k) Does <legend_label> appear as one of the legend labels in the graph ?

(l) What is the label or title of the X-axis ?

(m) What is the label or title of the Y-axis ?

(n) In how many cases, is the number of <figure_type> for a given <X_label> not equal to the number of legend labels ?

(o) What is the <Y_value> in/of < i^{th} X_tick> ?

(p) What is the <Y_value> of the i^{th} <legend_label> in < i^{th} X_tick> ?

(q) Does the <Y_label> monotonically increase over the <plural form of X_label> ?

(r) Is the <Y_label> of/in <legend_label1> strictly greater than the <Y_label> of/in <legend_label2> over the <plural form of X_label> ?

(s) Is the <Y_label> of/in <legend_label1> strictly less than the <Y_label> of/in <legend_label2> over the <plural form of X_label> ?

3. **Numeric Reasoning** :

(a) Across all <plural form of X_label>, what is the maximum <Y_label> ?

(b) Across all <plural form of X_label>, what is the minimum <Y_label> ?

(c) In which <X_label> was the <Y_label> maximum ?

(d) In which <X_label> was the <Y_label> minimum ?

(e) Across all <plural form of X_label>, what is the maximum <Y_label> of/in <legend_label> ?

(f) Across all <plural form of X_label>, what is the minimum <Y_label> of/in <legend_label> ?

(g) In which <singular form of X_label> was the <Y_label> of/in <legend_label> maximum ?

(h) In which <singular form of X_label> was the <Y_label> of/in <legend_label> minimum ?

(i) What is the sum of <title> ?

(j) What is the difference between the <Y_label> in < i^{th}x_tick> and < j^{th}x_tick> ?

(k) What is the average <Y_label> per <singular form of X_label> ?

(l) What is the median <Y_label> ?

(m) What is the total <Y_label> of/in <legend_label> in the graph?

(n) What is the difference between the <Y_label> of/in <legend_label> in < i^{th}x_tick> and that in < j^{th}x_tick> ?

(o) What is the difference between the <Y_label> of/in <legend_label1> in < i^{th}x_tick> and the <Y_label> of/in <legend_label2> in < j^{th}x_tick> ?

(p) What is the average <Y_label> of/in <legend_label> per <singular form of X_label> ?

(q) In the year < i^{th}x_tick>, what is the difference between the <Y_label> of/in <legend_label1> and <Y_label> of/in <legend_label2> ?

(r) What is the difference between the <Y_label> of/in <legend_label1> and <Y_label> of/in <legend_label2> in < i^{th}x_tick> ?

4. **Comparative Reasoning** :

(a) In how many <plural form of X_label>, is the <Y_label> greater than <N> units ?

(b) Do a majority of the <plural form of X_label> between < i^{th} x_tick> and < j^{th} x_tick¿ (inclusive/exclusive) have <Y_label> greater than N <units> ?

(c) What is the ratio of the <Y_label> in < i^{th} x_tick> to that in < j^{th} x_tick> ?

(d) Is the <Y_label> in < i^{th} x_tick> less than that in < j^{th} x_tick> ?

(e) In how many <plural form of X_label>, is the <Y_label> of/in <legend_label> greater than <N> <units>?

(f) What is the ratio of the <Y_label> of/in <legend_label1> in < i^{th} x_tick> to that in < j^{th}x_tick>?

(g) Is the <Y_label> of/in <legend_label> in < i^{th} x_tick> less than that in < j^{th} x_tick> ?

5. **Compound** :

(a) Is the difference between the <Y_label> in < i^{th}x_tick> and < j^{th}x_tick> greater than the difference between any two <plural form of X_label> ?

(b) What is the difference between the highest and the second highest <Y_label> ?

(c) Is the sum of the <Y_label> in < i^{th}x_tick> and < $(i+1)^{th}$x_tick> greater than the maximum <Y_label> across all <plural form of X_label> ?

(d) What is the difference between the highest and the lowest <Y_label> ?

(e) In how many <plural form of X_label>, is the <Y_label> greater than the average <Y_label> taken over all <plural form of X_label> ?

(f) Is the difference between the <Y_label> of/in <legend_label1> in < i^{th}x_tick> and < j^{th}x_tick> greater than the difference between the <Y_label> of/in <legend_label2> in < i^{th}x_tick> and < j^{th}x_tick> ?

(g) What is the difference between the highest and the second highest <Y_label> of/in <legend_label> ?

(h) What is the difference between the highest and the lowest <Y_label> of/in <legend_label> ?

(i) In how many <plural form of X_label>, is the <Y_label> of/in <legend_label> greater than the average <Y_label> of/in <legend_label> taken over all <plural form of X_label> ?

(j) Is it the case that in every <singular form of X_label>, the sum of the <Y_label> of/in <legend_label1> and <legend_label2> is greater than the <Y_label> of/in <legend_label3> ?

(k) Is the sum of the <Y_label> of/in <legend_label1> in < i^{th}x_tick> and < j^{th}x_tick> greater than the maximum <Y_label> of/in <legend_label2> across all <plural form of X_label>?

(l) Is it the case that in every <singular form of X_label>, the sum of the <Y_label> of/in <legend_label1> and <legend_label2> is greater than the sum of <Y_label> of <legend_label3> and <Y_label> of <legend_label4> ?

D. Question Types

Table 7 presents the distribution of the number of questions, categorized by their template type, that are present in each of the data splits.

Table 7: Detailed statistics of the dataset for different splits for different question types

Dataset Split	Question Type	#Questions	#Unique Answers
Train	Structural	1,743,760	34
	Data Retrieval	1,376,963	134,727
	Numeric Reasoning	1,280,736	339,508
	Comparative Reasoning	441,931	101,102
	Compound	890,503	164,030
Validation	Structural	373,568	34
	Data Retrieval	294,881	37,396
	Numeric Reasoning	274,339	77,687
	Comparative Reasoning	94,720	23,955
	Compound	190,960	41,408
Test-Familiar	Structural	373,646	34
	Data Retrieval	294,963	37,178
	Numeric Reasoning	274,218	77,627
	Comparative Reasoning	94,611	23,953
	Compound	190,875	41,286
Test-Novel	Structural	395,564	27
	Data Retrieval	372,439	9,899
	Numeric Reasoning	386,058	52,509
	Comparative Reasoning	170,236	19,674
	Compound	268,841	15,317

E. Answer Distribution

The answer distribution of the questions in the DIP dataset can be summarized in Table 8.

Table 8: Distribution of answers in the DIP dataset

Dataset Split	Type of Answers			
	Yes	No	Numeric	Text
Train	12.04%	15.32%	48.34%	24.30%
Validation	12.07%	15.30%	48.30%	24.33%
Test-Familiar	12.03%	15.34%	48.34%	24.29%
Test-Novel	11.58%	16.10%	51.73%	20.59%