IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao,
Wei Zhang, Zhou Yu, Xiaodan Liang, Song-Chun Zhu
1Center for Vision, Cognition, Learning and Autonomy, UCLA
2Sun Yat-sen University, 3East China Normal University, 4Columbia University

Abstract

Current visual question answering (VQA) tasks mainly consider answering human-annotated questions for natural images. However, aside from natural images, abstract diagrams with semantic richness are still understudied in visual understanding and reasoning research. In this work, we introduce a new challenge of Icon Question Answering (IconQA) with the goal of answering a question in an icon image context. We release IconQA, a large-scale dataset that consists of 107,439 questions and three sub-tasks: multi-image-choice, multi-text-choice, and filling-in-the-blank. The IconQA dataset is inspired by real-world diagram word problems that highlight the importance of abstract diagram understanding and comprehensive cognitive reasoning. Thus, IconQA requires not only perception skills like object recognition and text understanding, but also diverse cognitive reasoning skills, such as geometric reasoning, commonsense reasoning, and arithmetic reasoning. To facilitate potential IconQA models to learn semantic representations for icon images, we further release an icon dataset Icon645 which contains 645,687 colored icons on 377 classes. We conduct extensive user studies and blind experiments and reproduce a wide range of advanced VQA methods to benchmark the IconQA task. Also, we develop a strong IconQA baseline Patch-TRM that applies a pyramid cross-modal Transformer with input diagram embeddings pre-trained on the icon dataset. IconQA and Icon645 are available at https://iconqa.github.io.

1 Introduction

We are witnessing an exciting development of visual question answering (VQA) research in recent years. The long-standing goal of the VQA task is to exploit systems that can answer natural questions that correspond to visual information. Several datasets have been released to evaluate the systems’ visual and textual content understanding abilities [3, 57, 14, 21, 18, 52]. One of the underlying limitations of current VQA datasets is that they are focusing on answering visual questions for natural images. However, aside from natural pictures, abstract diagrams with visual and semantic richness account for a large proportion of the visual world. For instance, it is shown that emojis can express rich human sentiments [26, 10], and diagrams like icons can map the physical worlds into symbolic and aesthetic representations [31, 40, 24].

Some pioneering works attempt to propose datasets that are capable of answering questions for abstract diagrams. However, these datasets either address domain-specific charts, plots, and illustrations [26, 22], or are generated from limited templates [55, 48, 21]. These limitations impede their practical applications in real-world scenarios. For example, in elementary school, abstract diagrams in math world problems are involved with diverse objects and various reasoning skills [25].

To address these shortcomings, we introduce Icon Question Answering (IconQA), a new challenge for abstract diagram visual reasoning and question answering. The task, stemming from math word
problems for children [41], exhibits a promising potential to develop education assistants. We name the proposed task as IconQA because the images depict icons, which simplify recognition and allow us to focus on reasoning skills for further research. We release IconQA, a large-scale dataset that contains 107,439 QA pairs and covers three different sub-tasks: multiple-image-choice, multiple-text-choice and filling-in-the-blank. A typical IconQA problem is provided with an icon image and a question, and the answer is in the form of either a short piece of text or a choice from multiple visual or textual choices. Correctly answering IconQA questions needs diverse human intelligence skills. As the examples in Figure 1 show, IconQA poses new challenges for abstract diagram understanding like recognizing objects and identifying attributes. Besides, it is critical to develop diverse cognitive reasoning skills, including counting objects, comparing attributes, performing arithmetic operations, making logical inferences, completing spatial reasoning, or leveraging external commonsense to answer IconQA questions. More examples from the dataset are shown in Appendix A.1.

We use the IconQA dataset to benchmark various VQA approaches in the IconQA task, including four attention-based multimodal pooling methods [2, 28, 54, 11] and four Transformer-based pre-trained methods [33, 6, 53, 29], as illustrated in Figure 6. Also, we conduct extensive user studies to evaluate the performance differences between the algorithms and human beings. Three blind studies show that the IconQA dataset is robust against biased shortcuts when answering icon questions. We further develop a strong baseline called pyramid patch cross-modal Transformer (Patch-TRM), which effectively learns implicit visual and linguistic relationships in IconQA. Patch-TRM parses the diagrams into patch sequences in a spatial pyramid structure and learns a joint embeddings within a multimodal Transformer. Along with the IconQA dataset, we collect an auxiliary icon dataset, Icon645, that features 645,687 colored icons on 377 object classes. The icon dataset is used to pre-train the diagram embedding module in Patch-TRM to enhance abstract diagram understanding.

Our contributions can be summarized as 1) we propose a new challenge, IconQA, that requires abstract diagram understanding of icon images and diverse visual reasoning skills; 2) we establish two large-scale datasets: IconQA, a question answering dataset in the icon domain, and Icon645, an icon dataset for model pre-training; 3) we benchmark the IconQA dataset extensively via experiments on eight existing methods and develop a strong multimodal Transformer-based baseline.

2 Related Works

VQA Datasets. There have been efforts to develop datasets for the visual question answering (VQA) task since the first large-scale benchmark was introduced in [3]. Early released datasets [14, 30, 47, 52] contain natural images and related questions, where understanding the visual and textual contents is essential for question answering. Some recent datasets introduce questions that involve more diverse visual scenes or require external knowledge to answer, which leads to more complex visual and semantic reasoning for question answering. For example, CLEVR [21] is a synthetic dataset that serves as a diagnostic test for a range of visual reasoning abilities over combinations of three object
shapes. However, these datasets are limited to the natural image domain and pay little attention to abstract diagrams, which also have informative semantics and wide applications.

**Diagram QA Datasets.** To address the need for vision-and-language reasoning for diagrams, several abstract diagram QA datasets have been developed. For example, abstract VQA [3, 55] considers the task of answering questions on abstract scenes. Similarly, NLVR [48], FigureQA [23], and DVQA [22] feature diagrams of scientific plots that are generated with several figure types and question templates. However, questions and diagrams in these datasets are generated from limited templates, leading to the existence of unintended linguistic shortcuts for question answering. Some more works have proposed datasets of middle school math or science problems in more practical and complex scenarios [46, 27, 43, 44, 37]. A central limitation of the subject QA datasets is that they require complex domain-specific knowledge, which makes disentangling visual reasoning and domain knowledge difficult. Herein, we address these limitations by introducing the IconQA dataset, where only elementary commonsense is required. Through IconQA, we aim to provide a new benchmark for abstract scene understanding and learning different visual reasoning skills in real-world scenarios.

**VQA Methods.** Early VQA approaches usually combine multi-modal inputs by applying attention mechanisms over image regions or question words [28, 39, 38, 12, 54, 11]. Inspired by the semantic nature of VQA images, a line of approaches adopt object proposals from pre-trained object detectors and learn their semantic relationships [28, 54, 11]. As Transformers achieve excellent performance on vision tasks, pioneering works have attempted to use pre-trained models to learn visual representations for natural images in the VQA task [36, 33, 6, 29] and achieve significant improvements. However, current VQA models are not capable of extracting meaningful visual representations from abstract diagrams, as they require image embeddings or object proposals learned from natural images. Instead, we develop a strong baseline that feeds spatial patch sequences into a Transformer encoder that is powered by the embedding module pre-trained on our Icon645 dataset.

## 3 The IconQA Dataset

The IconQA dataset provides diverse questions that require abstract diagram recognition, comprehensive visual reasoning skills, and basic commonsense knowledge. IconQA consists of 107,439 questions split across three different sub-tasks. To the best of our knowledge, IconQA is the largest VQA dataset that focuses on real-world problems with icon images while involving multiple human intelligence reasoning abilities (see Table 4).

### 3.1 Data Collection

We aim to collect icon-based question answering pairs that involve multiple reasoning skills, such as visual reasoning and commonsense reasoning. To construct the IconQA dataset, which stems from real-world math word problems, we search for open-source math textbooks with rich icon images and diverse topics. Of those, we choose IXL Math Learning which compiles popular textbooks aligned to California Common Core Content Standards1. We ask well-trained crowd workers to collect problems that cover content from pre-K to third grade, as these problems usually contain abstract images and involve little to none complex domain knowledge. With the driven interest of visual reasoning over abstract images, we filter out the questions that do not accompany icon images or only have images in black and white. Redundant or repetitive data instances are also removed. Question choices are randomly shuffled to ensure a balanced answer distribution. See Appendix A for full details of the dataset collection and usage.

### 3.2 Data Analysis

Finally, we collect 107,439 IconQA data instances, where each data point contains a colored icon image, a natural language question, optional image or text choices, as well as a correct answer. The IconQA dataset consists of 107,439 questions and is divided into train, validation, and test splits with a ratio of 6:2:2, as shown in Table 1. The dataset consists of three sub-tasks: multi-image-choice, multi-text-choice, and filling-in-the-blank. The multi-image-choice sub-task is defined as choosing the correct image from a list of image candidates based on a given diagram and its corresponding question. Similarly, the multi-text-choice sub-task is defined as a multiple choice question with 2-5

1https://www.ixl.com/standards/california/math
text choices and an abstract diagram. The filling-in-the-blank sub-task is similar to the common VQA task, requiring a brief text answer for each question, except in IconQA, the images are icon images instead of natural images.

**Questions.** Figure 2 (a) illustrates the distribution of question lengths of each sub-task in the IconQA dataset. For simplicity, all questions longer than 35 words are counted as having 35 words. Questions in the multi-text-choice sub-task distribute more evenly, while for multi-img-choice, there is a long-tail distribution due to the complexity of textual scenarios. We find that some icon objects are frequently mentioned in the questions. In Figure 2 (b), the frequencies of the 40 most frequently mentioned icons are shown. These icon entities cover different daily-life objects such as animals, plants, shapes, food, etc. We cluster question sentences into different types based on frequent trigram prefixes starting the sentences. The distribution of questions is visualized in Figure 3. Importantly, the diversity in the question distribution implies the requirement of high-level understanding of textual and visual contents in IconQA. Figure 4 shows the word cloud of the question text in IconQA after eliminating the stop words. The most frequent words: shape, many, and object indicate that answering IconQA questions requires the model to identify a variety of geometric shapes and icon objects. Inspired by this, learning informative representations for icon images plays an important role in visual reasoning for the IconQA task.

**Skill Categories.** Our IconQA dataset contains questions of multiple different cognitive reasoning and arithmetic reasoning types that can be grouped into 13 categories, shown in Table 3. We annotate each question in IconQA with its corresponding skill types based on the tags provided by the original problem sources. Figure 5 shows the distributions of questions related to each skill. For instance, to answer 13.8% of the questions in IconQA, the model has to be capable of comparing object attributes. Additionally, each question can be related to up to three skills out of these 13 categories,
and on average, a question requires 1.63 skills. The detailed statistics are demonstrated in Table 2. In general, the filling-in-the-blank sub-task consists of questions that require the most number of skills, averaging 1.81 skills per question. 9.25% of the filling-in-the-blank questions require 3 skills. As the examples from IconQA shown in Figure 1, the first and second questions require the skills of scene understanding and spatial reasoning. The third example asks how many sticks exist in the diagram, requiring the basic ability of counting and basic algebra operations. As stated before, the IconQA dataset requires a wide range of skills for a model to perform well on IconQA.

<table>
<thead>
<tr>
<th>Skill types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry</td>
<td>Identify shapes, symmetry, transformations</td>
</tr>
<tr>
<td>Counting</td>
<td>Count objects, shapes</td>
</tr>
<tr>
<td>Comparing</td>
<td>Compare object attributes</td>
</tr>
<tr>
<td>Spatial</td>
<td>Identify spatial positions and relations</td>
</tr>
<tr>
<td>Scene</td>
<td>Understand abstract scenes</td>
</tr>
<tr>
<td>Pattern</td>
<td>Identify next and different patterns</td>
</tr>
<tr>
<td>Time</td>
<td>Identify time of clocks, events</td>
</tr>
<tr>
<td>Fraction</td>
<td>Perform fraction operations</td>
</tr>
<tr>
<td>Estimation</td>
<td>Estimate lengths, large numbers</td>
</tr>
<tr>
<td>Algebra</td>
<td>Perform algebraic operations</td>
</tr>
<tr>
<td>Measurement</td>
<td>Measure widths, lengths, heights</td>
</tr>
<tr>
<td>Commonsense</td>
<td>Apply external knowledge</td>
</tr>
<tr>
<td>Probability</td>
<td>Perform probability and statistics operations</td>
</tr>
</tbody>
</table>

Table 3: Definition of reasoning skill types.

Comparisons to Other Datasets. We compare our IconQA dataset with two datasets on natural images and five datasets on abstract diagrams in Table 4. To summarize, IconQA is different from these datasets in various aspects. Unlike natural images (VQA [3], CLEVR [21]) or abstract diagrams like scenes, charts, plots, and illustrations (VQA-Abstract [3], DVQA [22], NLVR [48], AI2D [26], Geometry3K [37]), IconQA features icon images and covers the largest object set of 388 classes. As questions in IconQA stem from real-world math problems and they may describe complex problem scenarios, IconQA has the longest question length among all related datasets. Furthermore, IconQA requires both commonsense and arithmetic reasoning due to its origin from real-world problems. Lastly, IconQA contains more QA task types including answering questions with image choices.

Table 4: Statistics for the IconQA dataset and comparisons with existing datasets.

3.3 Impact and Ethics

Impact & Usage. IconQA is useful for not only follow-up research projects but also real-world applications (e.g. K-6 education applications like tutoring assistants). Moreover, visual recognition in the abstract domain is essential to general AI agents, but rarely investigated in the community, posing new challenges in abstract and symbolic visual reasoning – a natural ability of human.

Social Ethics. Unlike VQA datasets in the natural image domain, IconQA is completely built upon abstract icon images. Therefore, it is less likely to be used in surveillance systems that might infringe on people’s privacy. Moreover, due to the abstract nature of the dataset, IconQA does not contain any sensitive personal information such as gender and race, nor does it contain data that might exacerbate biases against under-represented communities. Therefore, after careful examinations of our dataset, we think the dataset is unlikely to be used to harm people directly.

4 The Icon645 Dataset

As discussed in Section 3.2, IconQA questions are accompanied by abstract diagrams that cover a wide range of icon objects. Using existing backbone networks to extract image representations for
In this section, we first develop a patch cross-modal Transformer model (Patch-TRM) as a strong baseline for the IconQA task. To benchmark the IconQA dataset, we consider multi-modal pooling methods with attention mechanisms [2, 28, 11, 54], Transformer-based VQA approaches [36, 6, 53, 29], and three blind study methods as benchmark models, as summarized in Figure 6. Additionally, a user study is conducted to explore the performances of human beings in different age groups. In the sections below, we discuss the main principles of the core networks in the benchmarks we performed.

### 5.1 Our Baseline Model

Inspired by recent advances Transformer has achieved in vision-language tasks [33, 36], we develop a cross-modal Transformer model Patch-TRM for icon question answering. Taking the multi-image choice sub-task as an example, the overall architecture is shown in Figure 7. The diagram is first parsed into ordered patches in a hierarchical pyramid layout. These patches are then encoded by a pre-trained ResNet and passed through a vision Transformer. Question text is encoded by a language Transformer and fused with patch embeddings via the attention mechanism. The encoded image choices are concatenated with the joint diagram-question representation and then fed to a

---

**Table 5: Collected icon examples in the Icon645 dataset.**

<table>
<thead>
<tr>
<th>Icons</th>
<th>Examples</th>
<th>Icons</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bed</td>
<td>![Bed Icon]</td>
<td>Bucket</td>
<td>![Bucket Icon]</td>
</tr>
<tr>
<td>Cake</td>
<td>![Cake Icon]</td>
<td>Car</td>
<td>![Car Icon]</td>
</tr>
<tr>
<td>Castle</td>
<td>![Castle Icon]</td>
<td>Dog</td>
<td>![Dog Icon]</td>
</tr>
<tr>
<td>Giraffe</td>
<td>![Giraffe Icon]</td>
<td>Kite</td>
<td>![Kite Icon]</td>
</tr>
<tr>
<td>Soda</td>
<td>![Soda Icon]</td>
<td>Tree</td>
<td>![Tree Icon]</td>
</tr>
</tbody>
</table>

---

See Appendix B for the details of data collection and analysis.

---
classifier for question answering. The other two sub-tasks utilize similar network architectures, except that in the multi-text-choice sub-task, we use an LSTM encoder [17] for choice embedding, while filling-in-the-blank does not need a choice encoder.

Current dominant VQA methods either rely heavily on the ResNet backbone network to extract image features or depend on the Transformer encoders to learn image embeddings. However, these networks are pre-trained on natural images and are likely to fail to extract meaningful representations or reasonable object proposals when processing the diagrams in IconQA. Instead, we pre-train the ResNet network on the icon classification task with the icon dataset we compiled (Section 4). Patch-TRM hierarchically parses the diagram into patches that retain complete objects to a large extent, and the parsed patches are embedded by the pre-trained ResNet network before being fed into the vision Transformer. The hierarchical parsing structure, along with the ResNet pre-trained on icon data facilitate our Patch-TRM to learn informative diagram representations for the IconQA task. More details of the pre-training task are discussed in Section 6.4.

5.2 Benchmark Methods

Attention models. We construct four attention models for benchmarking. The first model implements Top-Down attention [2] for VQA, which is a strong attention method that applies free-form based attention on image representations from a pre-trained ResNet-101 network. The remaining three models utilize the bottom-up attention mechanism with the help of object detection proposals from Faster-RCNN [42]. Specifically, BAN [28] proposes a method that utilizes bilinear attention distributions to learn joint vision-language information. DFAF [11] is an advanced model that applies self-attention and cross-modal attention and introduces the information flow to help the model focus on target question words and image regions. The last approach, MCAN [54], learns the self-attention on the questions and images and the question-guided-attention of images jointly.

Transformer models. Four Transformer-based models are also implemented as benchmarks. ViLBERT [36] and UNITER [6] are two Transformer-based approaches that take image object proposals from Faster-RCNN [42] and question tokens as inputs. Specifically, ViLBERT learns the joint representation of the image content and the natural language content from image proposal regions and question tokens, while UNITER processes multimodal inputs simultaneously for joint visual and textual understanding. The last two benchmarks ViL [53] and ViLT [29] are more recently proposed Transformer models that take image patch tokens instead of object proposals as inputs when representing the image.

Blind study models. We develop three models to check for possible data biases in the IconQA dataset. A random baseline picks up one from the given choice candidates for the multiple-choice sub-tasks while predicts the answer by randomly selecting one from all possible answers in the train data for the filling-in-the-blank sub-task. Q-Only is set up similar to the Top-Down [2] model, but it only considers textual inputs. This baseline learns the question bias in the training set. I-Only also has a Top-Down architecture, but it only takes abstract diagrams as inputs, and tests the distribution biases in the images and answers in IconQA.

User study. To assess human performances in the IconQA task, we post the test set of IconQA on Amazon Mechanical Turk (AMT) and ask people to provide answers to the questions in the test set. We also ask the participants to provide us with their age group anonymously. We strongly encourage parents who have young children to let their children complete the questionnaires, as their answers...
Figure 8: Performance of humans in different age groups for the IconQA task. **Left**: Accuracy over three sub-tasks; **Right**: Accuracy over thirteen reasoning skills.

Table 6: Results on the IconQA dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sub-tasks (3)</th>
<th>Reasoning skills (13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>95.69</td>
<td>93.91</td>
</tr>
<tr>
<td>Random</td>
<td>41.70</td>
<td>36.87</td>
</tr>
<tr>
<td>Q-Only</td>
<td>41.64</td>
<td>36.86</td>
</tr>
<tr>
<td>I-Only</td>
<td>41.56</td>
<td>36.02</td>
</tr>
<tr>
<td>Top-Down [2]</td>
<td>75.92</td>
<td>68.51</td>
</tr>
<tr>
<td>BAN [23]</td>
<td>76.33</td>
<td>70.82</td>
</tr>
<tr>
<td>ViLBERT [33]</td>
<td>76.66</td>
<td>70.47</td>
</tr>
<tr>
<td>MCAN [54]</td>
<td>77.36</td>
<td>71.25</td>
</tr>
<tr>
<td>DEAF [11]</td>
<td>77.72</td>
<td>72.17</td>
</tr>
<tr>
<td>UNITER [6]</td>
<td>78.71</td>
<td>72.39</td>
</tr>
<tr>
<td>ViLT [29]</td>
<td>79.15</td>
<td>72.34</td>
</tr>
<tr>
<td>Patch-TRM (Ours)</td>
<td>82.66</td>
<td>75.19</td>
</tr>
</tbody>
</table>

Table 6: Results on the IconQA dataset.

6 Experiments

6.1 Training Details

Following prior work [3], all the baselines are trained on the IconQA training set, tuned on the validation set, and finally evaluated on the test set. Similar to [3], we choose accuracy as the evaluation metric. For the two multi-choice sub-tasks, the answer is regarded as correct only if it matches the ground truth. On the other hand, as the collected answers for filling-in-blank are short numbers, correct answers are expanded to include both the digital number and its corresponding words. More details of the benchmark setups and implementations can be found in Appendix E.1.

Our benchmarks and baselines are implemented using PyTorch. All experiments are run on one Nvidia RTX 3090 GPU. We use the Adamax optimizer with optimal learning rates of $7 \times 10^{-4}$, $8 \times 10^{-4}$, and $2 \times 10^{-3}$ on the three sub-tasks respectively. We apply a binary cross-entropy loss to train the multi-class classifier with a batch size of 64 and a maximum epoch of 50. The early stopping strategy is used when the validation accuracy stops improving for five consecutive epochs. It takes about 50, 30, and 10 minutes to train our baseline Patch-TRM on three sub-tasks respectively.

6.2 Experimental Results

Table 6 demonstrates the results of the benchmark methods and our baseline on the IconQA test set. The first three columns of the results represent the three sub-tasks: **multi-image-choice**, **multi-text-choice**, and **filling-in-the-blank** respectively. The remaining 13 columns illustrate the results of these approaches over problems that require different reasoning skills, as defined in Table 3.

Human performance. Out of the 54,896 collected answers, 9,620 are made by young children from age 3 to 8, 19,040 are made by adolescents from age 9 to 18, and 26,236 are made by adults.
Q: Which object is next to the one shaped like a cube?

![Image of objects]  

Figure 9: Text-to-image attention visualization.

Table 7: Ablation study in IconQA.

<table>
<thead>
<tr>
<th>Method</th>
<th>Img.</th>
<th>Txt.</th>
<th>Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patch-TRM w/o pre</td>
<td>82.01</td>
<td>72.72</td>
<td>81.67</td>
</tr>
<tr>
<td>Patch-TRM w/o att</td>
<td>80.63</td>
<td>68.00</td>
<td>80.29</td>
</tr>
<tr>
<td>Patch-TRM w/o pos</td>
<td>81.27</td>
<td>64.98</td>
<td>80.68</td>
</tr>
<tr>
<td>Patch-TRM V-CLS</td>
<td>80.15</td>
<td>63.90</td>
<td>70.27</td>
</tr>
<tr>
<td>Pyramid 1+4+9+16</td>
<td>82.45</td>
<td>68.76</td>
<td>82.19</td>
</tr>
<tr>
<td>Pyramid 1+4+9</td>
<td>80.61</td>
<td>67.42</td>
<td>81.36</td>
</tr>
<tr>
<td>Full model</td>
<td>82.96</td>
<td>75.21</td>
<td>83.10</td>
</tr>
</tbody>
</table>

The human performance over the three sub-tasks and thirteen skills is illustrated in Figure 8. As expected, young children do not answer the questions as well as adolescents or adults, suggesting that most participants answer their ages correctly. Moreover, the result shows that humans perform more consistently on all sub-tasks compared to machine algorithms. Interestingly, humans are outperformed by models quite significantly in questions that require numerical reasoning skills like probability, measurement, and estimation.

Analysis by Task Types. Humans outperform all benchmarks consistently over the sub-tasks and most reasoning skills. There is still a large gap to fill for future research of abstract diagram understanding and visual reasoning on the icon domain. The results achieved in blind studies of Q-only and I-only are close to random, showing that the IconQA dataset is robust and reliable in distribution. Our proposed Patch-TRM baseline outperforms current state-of-the-art VQA models in all three sub-tasks. These improvements mainly come from two insights: pre-training ResNet on icon images and taking a hierarchical approach with attention mechanism.

Analysis by Reasoning Types. Similarly, the Patch-TRM baseline obtains better results than the benchmarks over most reasoning skill types. Interestingly, in some skills such as estimation, measurement, and probability, Patch-TRM performs better than average human beings. This implies that neural networks have a promising potential to develop the basic ability of mathematical reasoning.

Quantitative Analysis. We visualize one example with the cross-modal attention map generated by our baseline Patch-TRM in Figure 9. The visualized attention shows that our baseline is capable of attending to the corresponding patch regions with higher weights given the input question.

6.3 Ablation Study

To study the functions of individual components in our model, we conduct an ablation analysis. Table 7 presents the results of different simplifications of our full model, where each implementation is trained on the IconQA train set and tested on the validation set. Instead of ResNet101 pre-trained on the icon classification task, Patch-TRM w/o pre utilizes ResNet101 pre-trained on natural image data for patch feature extraction. The decreasing performance of 0.95-2.49% indicates that pre-training backbones on tasks within similar domains is critical to downstream tasks. The attention mechanism helps to combine the image and question representations and improves the model performance by up to 7% compared to using simple concatenation (denoted as Patch-TRM w/o att). Positional embeddings of the ordered diagram patches benefit the vision Transformer by enabling it to learn spatial relationships among the patches, compared to the baseline without position embeddings (Patch-TRM w/o pos). Patch-TRM V-CLS uses the output embedding of [CLS] token as the diagram feature instead, which leads to a drastic performance decline. We have also experimented with coarse-grained patch cropping (e.g., Pyramid 1+4+9+16 denotes 30 patches, Pyramid 1+4+9 denotes 14 patches), which results in a performance degradation of 0.51% to 7.79%.

6.4 Icon Classification for Pre-training

The Icon645 dataset is collected to pre-train the backbone network for patch feature extraction. The dataset has a long-tailed distribution, and thus we address the class-imbalanced issue following previous studies on specific loss functions such as CB loss [3], Focal loss [34], and LDAM loss [5]. The metric of Top-5 accuracy is used to evaluate different model setups and the evaluation results are summarized in Table 8.

Table 8: Results for icon classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Medium</th>
<th>Tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet32 [16] + Focal Loss [34]</td>
<td>32.80</td>
<td>51.59</td>
<td>36.51</td>
</tr>
</tbody>
</table>

The Icon645 dataset is collected to pre-train the backbone network for patch feature extraction. The dataset has a long-tailed distribution, and thus we address the class-imbalanced issue following previous studies on specific loss functions such as CB loss [3], Focal loss [34], and LDAM loss [5]. The metric of Top-5 accuracy is used to evaluate different model setups and the evaluation results are summarized in Table 8.
Following [35], to demonstrate performances on different data parts, we divide the dataset into three balanced clusters: Head, Medium, and Tail, corresponding to 132, 122, and 123 classes respectively. All classes in Head have at least 1,000 instances, all classes in Medium have 300 - 999 instances, and all classes in Tail have fewer than 300 instances. As the results show, the backbone network ResNet101 with a re-balanced LDAM loss function achieves the best result for icon classification on Icon645. Consequently, we adopt this pre-trained ResNet101 network to extract patch features in our baseline Patch-TRM for IconQA.

7 Conclusion

In this work, we introduce IconQA, an open-source dataset of icon question answering in real-world scenarios for assessing the abilities of abstract diagram understanding and visual language reasoning. IconQA features 107,439 questions, three sub-tasks, and thirteen types of cognitive reasoning skills. We benchmark the IconQA task extensively with a user study, three blind studies, as well as multiple existing attention-based and Transformer-based approaches. We further develop a strong baseline, Patch-TRM, which parses the diagram in a pyramid layout and applies cross-modal Transformers with attention mechanism to learn the meaningful joint diagram-question feature. Additionally, we introduce Icon645, a large-scale icon dataset that is useful to pre-train the diagram encoding network used in Patch-TRM for the IconQA task.

By releasing a new dataset of icon question answering for abstract diagram understanding and visual language reasoning, we envision that IconQA will facilitate a wide range of research in computer vision and natural language processing, as well as smart education applications like tutoring systems, to invent the future of AI for science education.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Appendix A.13.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Section 3.3.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see our project page at https://iconqa.github.io.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Section 6.1 for training details. For details on benchmark model settings, see Appendix E.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Section 6.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See https://github.com/lupantech/IconQA#license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All datasets are available on the IconQA website https://iconqa.github.io, or the github repository https://github.com/lupantech/IconQA.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] As we discuss in Section 3.3, our datasets do not contain identifiable or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix D.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix D.3.
A The IconQA Dataset

The following datasheet follows the format suggested in this paper [13].

A.1 More Examples

![Image of examples]

Figure 10: More examples in the IconQA dataset. Top: The multi-image-choice sub-task. Middle: The multi-text-choice sub-task. Bottom: The filling-in-the-blank sub-task.

A.2 Dataset Label

The IconQA dataset label is shown in Figure 11.

A.3 Question Skill Categories

The questions we collected contain meta-information including question topics, chapter names, image names, etc. After extensive data exploration by well-informed individuals, we designed a set of rules that map each question to 1-3 of the 13 categories based on trigger words in metadata. The rules for trigger words are list in Table 9.

A.4 Links

The link to download the IconQA dataset can be found at iconqa.github.io.

IconQA Dataset Facts

<table>
<thead>
<tr>
<th>Metadata</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Website</td>
<td><a href="https://iconqa.github.io">https://iconqa.github.io</a></td>
</tr>
<tr>
<td>Instances</td>
<td>107,439</td>
</tr>
<tr>
<td>Format</td>
<td>.png, .json</td>
</tr>
<tr>
<td>License</td>
<td>CC BY-NC-SA</td>
</tr>
<tr>
<td>Original Use Case</td>
<td>Training VQA systems</td>
</tr>
</tbody>
</table>

| Composition  |  |
| Sample or Complete | Sample from ixl.com |
| Missing Data     | No data is missing |

| Collection |  |
| Sampling Strategy | See the main paper |
| Author Consent | None needed |

| Cleaning and Labeling |  |
| Cleaning Done | Repetitions and redundancies removed |
| Labeling Done | Yes |

| Uses and Distribution |  |
| Notable Uses | Training VQA systems |
| Original Distribution | Check dataset website |

| Maintenance and Evolution |  |
| Corrections or Erratum | None as of now |
| Methods to Extend | Contact the author |

<table>
<thead>
<tr>
<th>Breakdown</th>
<th>% of Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>multi-image-choice</td>
<td>57,672 items</td>
</tr>
<tr>
<td>multi-text-choice</td>
<td>31,578 items</td>
</tr>
<tr>
<td>fill-in-the-blank</td>
<td>18,189 items</td>
</tr>
</tbody>
</table>

Figure 11: IconQA dataset label, created with the template from the paper [4].

Table 9: Trigger words in metadata for skill categories.

<table>
<thead>
<tr>
<th>Skill types</th>
<th>Trigger words in metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry</td>
<td>name the shape, shapes of, classify shapes, solid, corners, faces, edges, vertices, sides, dimensional, rectangle, circle, triangle, square, rhombus, sphere, cylinder, cone, cubes, hexagon, perimeter, area, curved, open and close, flip turn, symmetry</td>
</tr>
<tr>
<td>Counting</td>
<td>count, tally, a group, ordinal number, area, even or odd, place value, represent numbers, comparing review, equal sides, square corners, one more, one less, fewer, enough, more.</td>
</tr>
<tr>
<td>Comparing</td>
<td>compare, comparing, more, less, fewer, enough, wide and narrow, light and heavy, long and short, tall and short, match analog and digital</td>
</tr>
<tr>
<td>Spatial</td>
<td>top, above, below, beside, next to, inside and outside, left</td>
</tr>
<tr>
<td>Scene</td>
<td>problems with pictures, beside, above, inside and outside, wide and narrow, objects</td>
</tr>
<tr>
<td>Pattern</td>
<td>the next, comes next, ordinal number, different</td>
</tr>
<tr>
<td>Time</td>
<td>clock, am or pm, elapsed time, times</td>
</tr>
<tr>
<td>Fraction</td>
<td>equal parts, halves, thirds, fourths, fraction</td>
</tr>
<tr>
<td>Estimation</td>
<td>estimate, measure</td>
</tr>
<tr>
<td>Algebra</td>
<td>count to fill, skip count, tally, even or odd, tens and ones, thousands, of ten, elapsed time, perimeter, area, divide</td>
</tr>
<tr>
<td>Measurement</td>
<td>measure</td>
</tr>
<tr>
<td>Commonsense</td>
<td>light and heavy, compare size, holds more or less, am or pm, times of, tool</td>
</tr>
<tr>
<td>Probability</td>
<td>likely</td>
</tr>
</tbody>
</table>

A.5 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  - IconQA is created to provide researchers with a wide range of VQA data on the abstract image domain. Currently, existing datasets 1) are limited to natural images, or 2) contain diagrams generated with templates, and therefore lack linguistic variation, or 3) include too much domain specific knowledge. We believe that no other abstract diagram QA dataset exists that covers such a wide range of perceptive and cognitive abilities without requiring complicated domain knowledge.
• Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  – This dataset was created under the combined effort of multiple researchers from University of California, Los Angeles, Sun Yat-sen University, East China Normal University, and Columbia University.

• Who funded the creation of the dataset?
  – The project received no funding or associated grant.

A.6 Composition

• What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
  – Each instance is a complete icon question answering task.

• How many instances are there in total (of each type, if appropriate)?
  – There are a total of 107,439 instances. 57,672 are multi-image-choice questions, 31,578 are multi-text-choice questions, and 18,189 are filling-in-the-blank questions.

• Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  – The dataset does not contain all possible instances.

• What data does each instance consist of?
  – Each instance in IconQA includes a textual question, an image, and multiple optional visual / textual choices. We also included some metadata about each question, such as the skill type, question type, etc.

• Is there a label or target associated with each instance?
  – Yes, each question is associated with a ground truth answer.

• Is any information missing from individual instances?
  – No. All related information is included in the dataset.

• Are there recommended data splits (e.g., training, development/validation, testing)?
  – Yes. Following conventions in the field, we have splitted the dataset into a training set, a validation set, and a test set with a 0.6:0.2:0.2 ratio.

• Are there any errors, sources of noise, or redundancies in the dataset?
  – We randomly selected 1,000 questions from each sub-task and ask an experienced expert to double check the answers carefully. Among the 1,000 multi-image-choice questions, only 1 error was found. Among the 2,000 questions of the other two sub-tasks, no error was found.
  – In the multi-image-choice sub-task, questions that ask “Which two are exactly the same?” might be a source of noise for certain use cases, as in the data label, only one correct answer out of the two is given. We intend to address this problem in the later versions of the dataset.

• Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  – The dataset is self-contained. All related information is included in the dataset.

• Does the dataset contain data that might be considered confidential?
  – No, the dataset does not contain anything related to any individual.

• Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  – No, the dataset does not contain anything offensive.
A.7 Collection Process

- How was the data associated with each instance acquired?
  - The data is publicly available on ixl.com. More details are included in the main paper.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
  - We implemented an integrated graphic user interface tool using Python to help crowd workers to collect the data.
- Over what timeframe was the data collected?
  - The dataset was finally completed in March, 2021 after three months of data collection, cleaning, and preprocessing.
- Were any ethical review processes conducted (e.g., by an institutional review board)?
  - No, we did not conduct an ethical review under the assumption that math and science questions designed for young children do not contain any discriminative or offensive content.
- Does the dataset relate to people?
  - No, the dataset does not relate to people.

A.8 Preprocessing

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - We cropped white space from each diagram in IconQA to tighten it up. Questions with invalid diagrams, answers, or choices were filtered out. Redundant instances were removed based on the metrics of exact question text matching and diagram similarity.
- Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  - Yes, each QA data is accompanied with reasoning skill types and a grade level for comprehensive analysis of different benchmarks.
- Is the software used to preprocess/clean/label the instances available?
  - The data preprocessing and cleaning was done using Python.

A.9 Use Cases

- Has the dataset been used for any tasks already?
  - Yes, we developed a baseline model of cross-modal Transformers and multiple benchmarks for icon question answering, and we trained the models on the IconQA dataset. For more details, refer to Section 5 of the main paper.
- Is there a repository that links to any or all papers or systems that use the dataset?
  - Yes, you can access the code to our model at github.com/lupantech/IconQA.
- What (other) tasks could the dataset be used for?
  - Currently, the dataset is intended for training visual question answering systems to access the abilities of diagram upstanding and visual reasoning. More uses could be explored in research of computer vision, natural language processing, and multimodal learning, as well as applications in smart education like tutoring systems.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  - No.
A.10 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  - The dataset is free to all under the condition that the dataset is used for non-commercial purposes only.

- How will the dataset be distributed (e.g., tarball on website, API, GitHub)?
  - You can find our dataset both on the IconQA website iconqa.github.io, or the github repository github.com/lupantech/IconQA

- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  - The dataset will be distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license.

- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  - The source of the data instances, IXL, does not allow the data to be used commercially.

A.11 Maintenance

- Who is supporting/hosting/maintaining the dataset?
  - The dataset is maintained by the paper’s authors.

- How can the owner/curator/manager of the dataset be contacted?
  - The contact information of the authors can be found at the beginning of the main paper.

- Is there an erratum?
  - Currently, little errors have been found in the dataset. However, if errors were to be found, an erratum will be included in the repository.

- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  - If the dataset were to be updated, all versions will be available on the dataset website.

- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  - Please contact the author through email.

A.12 Novelty

IconQA presents new challenges in icon understanding and cognitive reasoning to many existing visual reasoning methods. 1) Icons feature intrinsic natures of abstract symbolism, varied styles, and ambiguous semantics, which differs from natural images significantly. 2) Since there is a lack of high-quality annotation data for icon diagrams, it restricts current mainstream data-driven visual methods to transfer smoothly to the icon domain. 3) As 107,439 questions in IconQA stem from real-world math word problems, it has made 13 different cognitive reasoning skills essential, including spatial reasoning, commonsense reasoning, estimation, and arithmetic calculation.

A.13 Limitations and Future Work

Dataset Expansion. As discussed in Section 3, IconQA focuses on colored abstract diagrams and questions of third grade and below to simplify the context scenarios and attract the community’s attention on diagram understanding and visual reasoning. We would like to expand the dataset to provide greater diversity of diagram formats, grade levels, icon classes and reasoning skill types.

Fine-grained Annotations. IconQA benchmarks the visual question answering task in the icon domain and releases a dataset of questions, diagrams and answers. But it would be beneficial to include the object-level parsing annotations and textual explanations for each diagram and question, which facilitates future research on semantic diagram parsing and transparent visual reasoning.

3https://creativecommons.org/licenses/by-nc-sa/4.0
B The Icon645 Dataset

The following datasheet follows the format suggested in this paper [13].

B.1 Dataset Statistics

Table 10: Statistics for the Icon645 dataset.

<table>
<thead>
<tr>
<th>Data</th>
<th>#Classes</th>
<th>#Icons</th>
<th>Min Size</th>
<th>Max Size</th>
<th>Colored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Icon645</td>
<td>377</td>
<td>645,687</td>
<td>64×64</td>
<td>256×256</td>
<td>✓</td>
</tr>
</tbody>
</table>

B.2 Dataset Label

<table>
<thead>
<tr>
<th>Icon645 Dataset Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Website</td>
</tr>
<tr>
<td><a href="https://iconqa.github.io">https://iconqa.github.io</a></td>
</tr>
<tr>
<td>Metadata</td>
</tr>
<tr>
<td>Classes</td>
</tr>
<tr>
<td>Instances</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Image Sizes</td>
</tr>
<tr>
<td>License</td>
</tr>
<tr>
<td>Original Use Case</td>
</tr>
<tr>
<td>Composition</td>
</tr>
<tr>
<td>Sample or Complete</td>
</tr>
<tr>
<td>Missing Data</td>
</tr>
<tr>
<td>Collection</td>
</tr>
<tr>
<td>Sampling Strategy</td>
</tr>
<tr>
<td>Author Consent</td>
</tr>
<tr>
<td>Cleaning and Labeling</td>
</tr>
<tr>
<td>Cleaning Done</td>
</tr>
<tr>
<td>Labeling Done</td>
</tr>
<tr>
<td>Uses and Distribution</td>
</tr>
<tr>
<td>Notable Uses</td>
</tr>
<tr>
<td>Other Uses</td>
</tr>
<tr>
<td>Original Distribution</td>
</tr>
<tr>
<td>Maintenance and Evolution</td>
</tr>
<tr>
<td>Corrections or Erratum</td>
</tr>
<tr>
<td>Methods to Extend</td>
</tr>
</tbody>
</table>

Figure 12: Icon645 dataset label, created with the template from the paper [4].

B.3 Links

The link to download the IconQA dataset can be found on iconqa.github.io.

B.4 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  - Icon645 was created for the purpose of pre-training image encoders on the icon image classification task. Presently, no other dataset provides such a large variety of abstract icons with appropriate labels.

- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  - This dataset was created under the combined effort of multiple researchers from University of California, Los Angeles, Sun Yat-sen University, East China Normal University, and Columbia University.
• Who funded the creation of the dataset?
  – The project received no funding or associated grant.

B.5 Composition
• What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
  – Each instance is a single colored icon image with size between $64 \times 64$ and $256 \times 256$ pixels.
• How many instances are there in total (of each type, if appropriate)?
  – There are a total of 645,687 instances categorized into 377 classes.
• Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  – The dataset is a sample of the Flaticon library. Only 377 classes of icons that satisfy our requirements outlined in the paper are included in the dataset.
• What data does each instance consist of?
  – Each instance is a single icon image
• Is there a label or target associated with each instance?
  – Yes, Each image is given a text label, specifying its class.
• Is any information missing from individual instances?
  – No. All related information is included in the dataset.
• Are there recommended data splits (e.g., training, development/validation, testing)?
  – No. The user can decide how they want to split the dataset.
• Are there any errors, sources of noise, or redundancies in the dataset?
  – No.
• Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  – The dataset is self-contained. All related information is included in the dataset.
• Does the dataset contain data that might be considered confidential?
  – No, the dataset does not contain anything related to any individual.
• Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  – No, the dataset does not contain anything offensive.

B.6 Collection Process
• How was the data associated with each instance acquired?
  – The data is publicly available on flaticon.com. More details are included in the main paper.
• What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
  – We implemented a program to retrieve the target 377 icon classes using Python.
• Over what timeframe was the data collected?
  – The dataset was finally completed in March, 2021 after three months of data collection, cleaning and preprocessing.
• Does the dataset relate to people?
  – No, the dataset does not relate to people.
B.7 Preprocessing

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - We cropped white space from each icon diagram in Icon645 to tighten it up. Black and white icons were filtered out. Redundant instances were removed based on the metric of diagram similarity.
- Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  - No.
- Is the software used to preprocess/clean/label the instances available?
  - The data preprocessing and cleaning was done using Python.

B.8 Use Cases

- Has the dataset been used for any tasks already?
  - Yes, we have used the dataset to pre-train an abstract image encoder to act as the backbone network in our Patch-TRM model.
- Is there a repository that links to any or all papers or systems that use the dataset?
  - Yes, you can access the code to our model at github.com/lupantech/IconQA.
- What (other) tasks could the dataset be used for?
  - Currently, the dataset is intended for training abstract icon image classifiers. Other possibilities could be explored in the future.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  - No.

B.9 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  - The dataset is free to all under the condition that the dataset is used for non-commercial purposes only.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
  - The dataset will be accessible on github.com/lupantech/IconQA.
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  - The dataset will be distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license.
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  - The source of the data instances, Flaticon, does not allow the data to be used commercially.

B.10 Maintenance

- Who is supporting/hosting/maintaining the dataset?
  - The dataset is maintained by the paper’s authors.
- How can the owner/curator/manager of the dataset be contacted?
  - The contact information of the authors can be found at the beginning of the main paper.

4https://creativecommons.org/licenses/by-nc-sa/4.0
• Is there an erratum?
  – Currently, no error has been found in the dataset. However, if errors were to be found, an erratum will be included in the repository.

• Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  – If the dataset were to be updated, all versions will be available on the dataset website.

• If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  – Please contact the author through email.

C Details of Baseline Patch-TRM

We develop a patch cross-modal Transformer model (Patch-TRM) as a strong baseline for the IconQA task as illustrated in Figure 7. We will introduce the details of Patch-TRM as follows.

C.1 Diagram Encoder

Similar to natural images in most VQA datasets, abstract diagrams also have rich visual and semantic information that is critical to answering questions. Current dominant VQA methods [3, 2, 28, 11, 54, 19, 1] either extract high-level visual representations from a pre-trained ResNet backbone network [16] in a top-down fashion, or apply a bottom-up mechanism to extract semantic representations via an object detector, such as a model based on Faster R-CNN [42]. However, these methods depend heavily on the backbone network, which is pre-trained on natural images. When processing diagrams in IconQA, they are likely to fail to extract meaningful representations or reasonable object proposals. Inspired by the early progress in using hierarchical scene layout to parse images [32, 56, 51] and the recent advances in Transformer-based image encoding [36, 33, 53], we develop a method that splits diagrams into hierarchical patch sequences from a pyramid structure and learns their visual representations using a visual Transformer.

As diagrams in IconQA have more varied aspect ratios than natural images, we add blank paddings at the bottom or on the right side of the images to ensure that they are square-shaped. Each padded diagram is then cropped into a set of patch sequences with different scales. The padding operation and the hierarchical scene layout can facilitate extracting complete objects that retain specific semantics. Let \( p = [p_1, p_2, \ldots, p_n] \) denote the patch sequence in the splitting order from the original diagram. From each patch sequence, we extract the visual features using a ResNet model and represent the features as \( f_p = [f_{p_1}, f_{p_2}, \ldots, f_{p_n}] \). The representation for each patch, \( f_{p_i} \), is then summed up with its positional embedding with respect to its sequential index \( i \). Finally, the updated visual patch embeddings pass through a standard multi-layer Transformer [50] to learn high-level visual representations \( h_p = [h_{[CLS]}, h_{p_1}, h_{p_2}, \ldots, h_{p_n}] \). Here, the trainable token \([CLS]\), which is added to the Transformer inputs, learns the global meaning of these sequences. As mentioned before, it is not feasible to use existing pre-trained ResNet to process abstract diagrams due to a lack of similar resources for pre-training. So we pre-train the ResNet on icon classification with the icon dataset we compiled (Section 4). More details of the pre-training task are discussed in Section 6.4.

C.2 Language Encoder

Questions in IconQA have a wide distribution of question lengths, so we follow the recent approaches [50, 20, 49, 33, 36] that apply the BERT model [9] to embed question texts, rather than using traditional LSTM [17] or GRU [7] for long sequence encoding. Given the question \( w_0, w_1, \ldots, w_t \), the input is formatted as \([CLS], w_0, w_1, \ldots, w_t\) \). We use the WordPiece [45] subword tokenizer and the resulting sequence is padded to the maximum length. Similar to other methods that use BERT as sentence encoders, we consider the output corresponding to the first token \([CLS]\) as the embedding of the entire question, noted as \( h_q \).

C.3 Answer Reasoning

Given the image patch representation \( h_p \in \mathbb{R}^{n \times k} \), and question embedding \( h_q \in \mathbb{R}^k \), where \( n \) denotes the number of diagram patches and \( k \) denotes the learned embedding size of the patches, we
apply a cross-modal attention to learn their joint representation:

\[ a = \text{softmax}(W_p h_p \circ W_q h_q), \]  

\[ h_v = \sum_i a(i) \times h_{pi}, \]  

where \( W_p \) and \( W_q \) are learnable mapping parameters, and \( \circ \) is the element-wise product operator. The joint representation \( h_v \) is calculated as the weighted sum over all diagram patches.

Before predicing the answer, multiple choice candidates need to be encoded. Taking the multi-image-choice task as an example, each image choice is encoded as the output of the last pooling layer of the pre-trained ResNet. The encoded image choice is denoted as \( h_c \in \mathbb{R}^{m \times k} \), where \( m \) is the number of the candidates. The choice embeddings are concatenated with the diagram-question representation, and then the resulted embeddings are fed to a classifier over the candidates:

\[ p_{\text{ans}} = \text{softmax}(W_a ([h_v; h_c]) + b_a), \]  

where \( W_a \) and \( b_a \) are classifier parameters, and \( p_{\text{ans}} \) is the probability of the predicated answer choice.

Similarly, in the multi-text-choice sub-task, the answer is prediced over text choices, except that each text choice is embedded with LSTM layers first. We formulate the filling-in-the-blank sub-task as a multi-class classification problem from all possible answers in the training data, as most VQA works do. After generating the joint encoding for the input diagram and question, a linear classifier is trained to predict the final answer.

D User Study

D.1 Crowd Sourcing Method

Using Amazon Mechanical Turk (AMT), we ask people to provide answers to the questions in the test set along with their age group. We also strongly encourage parents who have young children to let their children complete the questionnaires, as their answers give us insights to how the designed audience of these questions perform. The test set is split into batches of 20 questions, which we call a task, with each task assigned to 3 crowd workers on AMT. This amounts to a total of 64,467 effective test set answers.

D.2 Quality Assurance

To ensure the truthfulness of the age information, we ask the participants to select their age at both the beginning and the end of the questionnaire, with the age choices appearing in 2 different orders.

To ensure the quality of the answers, we include 4 attention check questions: 3 of which are about the instructions, making sure that the participants read the instructions carefully. We also add an extra fake question in the middle for each choosing an image choice and choosing a text choice task, instructing them to choose the fourth choice despite what the choices are. Figure 13 shows the instructions and the first three attention check questions. Figure 14 shows the fake question along with the age confirmation. Figure 15, 16, and 17 are example questions for three sub-tasks respectively. We also make sure that the workers answering our tasks have a history HIT approval rate of at least 95% and a previous approval count of 1,000.

In summary, for each Human Intelligence Task (HIT) on AMT, we have 2 age questions, 4 attention check questions, and 20 real questions from the IconQA test set. Among the 64,467 test answers, we filter out 1) the questionnaires that do not pass the 4 attention check questions, 2) the questionnaires that do not answer consistently for the two age-related questions, 3) the questionnaires that are finished unreasonably slowly/quickly. After filtering, we have 54,896 effective question answers, which we believe is a decently large sample for the human performance study.

D.3 Worker Compensation

For each batch of multiple choice questions, we provide a monetary compensation of 10 US cents. For each batch of filling-in-the-blank questions, we provide a compensation of 20 cents due to the
increased difficulty. We decide upon these numbers after a few timed test trials run by ourselves. we
find that these numbers enable the workers to acquire 6 USD per hour, an above average hourly wage on the AMT platform [15]. The total spending in the end sums up to 452.52 USD.

E Experiments

E.1 Experimental Details

We use the same learning parameters set in Top-Down [2] when evaluating the eight benchmarks listed in Section 5 and our developed baseline Patch-TRM. Some crucial parameters used in our model are clarified below.

Our Baseline Model. For our baseline Patch-TRM, each diagram is split four times by varied scales, resulting in 79 (1+4+9+16+49) patches totally. After resizing them to 224 × 224, patch visual features are extracted from the last pooling layer, resulting in a 2048-d feature vector. The ResNet network used to embed the patches is pre-trained on the icon classification task as discussed in Section 6.4. The patch Transformer has one layer of Transformer block with four attention heads and outputs embeddings with a hidden state size of 768. A small pre-trained BERT model [49] is used to encode the question text in the language encoder.

Attention models. For Top-Down, the attention-based baselines use 7 × 7 × 2048-d features from the last convolution layer. For BAN [28], DFAF [11], and MCAN [54], image features of dimension 2,048 are extracted from Faster R-CNN [42]. Question words in these attention models are encoded into features of dimension 1,024 by GRU [7]. And the visual and textual features are then embedded into 1,024 dimensions with the corresponding attention mechanisms and fusion methods reported in original works.

Transformer models. For ViLBERT [36] and UNITER [6], we use Faster R-CNN [42] to extract 36 proposal regions as the visual inputs. Both ViL [53] and ViLT [29] use ViT-B/32 pre-trained on ImageNet to encode the image embeddings. The hidden size is set as 768, the layer depth is 32, and the input image is sliced into patches with a size of 32. For ViL, we use two dependent Transformers to embed the question and image respectively.
E.2 Human Performance

The detailed results for human performance in the IconQA task are shown in Table 11.

Table 11: Human performance in the IconQA task.

<table>
<thead>
<tr>
<th>Sub-tasks (3)</th>
<th>Reasoning skills (13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>95.69</td>
</tr>
<tr>
<td>Human (3-8)</td>
<td>94.58</td>
</tr>
<tr>
<td>Human (9-18)</td>
<td>94.63</td>
</tr>
<tr>
<td>Human (19+)</td>
<td>97.34</td>
</tr>
</tbody>
</table>

E.3 Quantitative Analysis

Figure 18 presents five examples from the IconQA test set predicted by our Patch-TRM baseline for each sub-task. Although Patch-TRM achieves promising results for most problems in IconQA, it still fails to address some complicated cases. For example, it encounters difficulties in identifying dense objects and making multi-hop reasoning.

Figure 18: Result examples predicted by our Patch-TRM model in the IconQA test set. Top: Multi-image-choice sub-task. Middle: Multi-text-choice sub-task. Bottom: Filling-in-the-blank sub-task. Correctly predicted answers are highlighted by green, while wrong ones are highlighted by red.