Chart2Table: A Neural Model for Converting Charts to a Latex Script of Table

Neelu Verma
Computer Science and Engineering
Indian Institute of Technology
Jodhpur, India
verma.14@iitj.ac.in

Arpit Gupta
Electrical Engineering
Indian Institute of Technology
Jodhpur, India
gupta.25@iitj.ac.in

Raghav Ranjan
Electrical Engineering
Indian Institute of Technology
Jodhpur, India
ranjan.1@iitj.ac.in

Anand Mishra
Computer Science and Engineering
Indian Institute of Technology
Jodhpur, India
mishra@iitj.ac.in

***Abstract*—This paper presents a novel Neural Network approach for converting charts to a latex script of the table. Most scientific documents contain charts like line charts, pie charts, bar charts, and other plots to visualize the results better and comprehensively. Moreover, these information-rich charts are hard to read and interpret by most OCR engines. Our first main contribution is a highly complex and real world dataset of chart images named ChartReal. Second, the Seq2Seq model processes the text and converts it to a table's latex script. The third contribution toward this research compares our work on synthetic and some existing datasets. We study the chart to latex script of the table by training several models, including the current Sequence to Sequence neural machine translation model as a primary baseline. Calculating the BLEU score on a synthetic dataset, the results are quite good as BLEU Scores Average is 48% on 5000 training data and 1000 new test data with Random string as labels, BLEU Scores Average is 40%**

***Keywords*-component; ChartReal, Seq2Seq, Computer Vision, Object detection, Information Retrieval, Chart data extraction.**

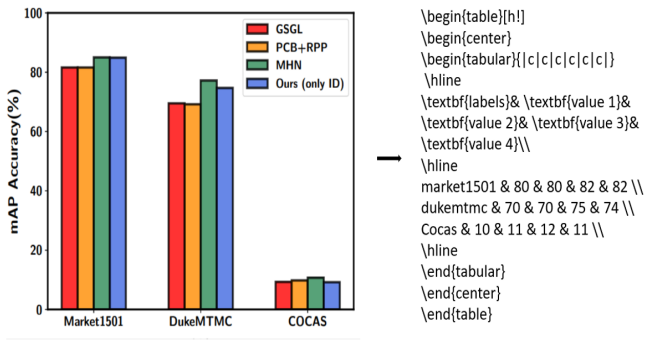

Figure 1. **Chart2Table**: The Overall objective of our research to show input image is converting into latex script of table. Left side represent the image from ChartReal dataset and right side show the corresponding latex script

I. Introduction

Optical character recognition (OCR) is one of the most successful applications of machine learning. Most OCR engines [1], [2] read the textual content with reasonably high performance in document images, but they tend to ignore the graphics contents by just tagging them. These graphic contents often contain information-rich charts such as bar, pie, and line charts in scientific documents. Our goal is to fill this gap in modern OCR engines by developing a method to make charts more interpretable. To this end, we propose a new task in document image literature, namely *chart2table*. This task's goal is illustrated in Figure 1 is to convert unstructured chart images to semi-structured and machine-readable tables (or their equivalent LaTex scripts).

The *chart2table* is a pathetically challenging task due to inherent complex structures of charts and diversity among them. Consider an example shown in Figure 1. It is non-trivial to precisely convert a bar chart shown on the left to a table that encodes all the chart information. The heuristic or rule-based methods, e.g., based on image processing techniques may work well for a category of chart image, but does not generalize well to a novel chart with an entirely different visualization, e.g., a chart without legend and having horizontal bars may require a different rule. We see an analogy between *chart2table* and image captioning problem [3], [4]. In the image captioning problem, given a natural scene image, the goal is to generate a natural language description. Whereas in *chart2table*, given a chart image, we aim to generate a LaTex script corresponding to a table that encodes all the information present in the chart image. This analogy motivates us to explore and adapt state-of-the-art image captioning techniques to our problem.

Sequence to sequence models has shown promising per-

formance in image captioning tasks. We adapt them for *chart2table* and propose the following solution where the model takes input as a sequence of bounding boxes and text labels and produce a latex script of the table.

.

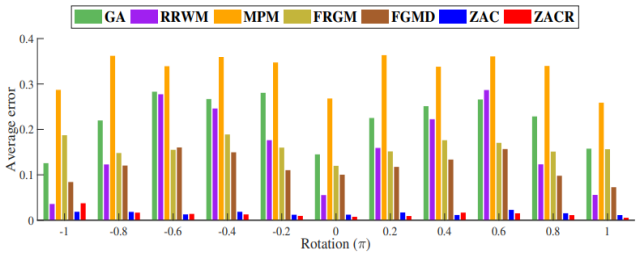

Figure 2. Sample image that is hard to annotate by existing methods

Contributions of this work are three folds,

1) We draw the document image analysis community's attention to the important but unexplored problem of interpreting chart images and making them better machine-readable by converting them to semi-structured table encoding all the information in the chart image. We refer to this new task as *chart2table*.

2) We introduce *ChartReal*: a first fully annotated dataset for evaluating the performance of *chart2table*. This dataset contains chart images collected from various technical reports, research papers, magazines, and newspapers. Further, each of these chart images is annotated by a human annotator to get the corresponding table.

3) We perform a systematic study of the problem and provide robust baseline methods. These proposed baseline methods are built on modern image captioning based techniques. Further, we evaluate the performance of the baseline methods in a principled way. We firmly believe the new task and associated dataset will enhance the interest in understanding chart images.

II. Literature Review

Any statistical charts contains complex, large information and a set of structural text and data. Charts/figures are impressive, effective and extensively used representations of data in any documents. It can contain various text information like title, axis-title, legend, axis-tick, and information of data encoded into width or height of bars. For this purpose one must firstly detect the bounding boxes of texts and bars. The system that automatically extracts information from charts would provide great benefits in knowledge management within the company because no one wants to look at formidable data. Such knowledge can be combined with other data sets to further enhance business value.

Several application uses the result of chart data extraction. It is progressively hard to find robust technologies that can understand and predict the documents and charts along with the ability to provide text labels. Large images synthetic dataset is publicly available but it is hard to find charts with annotations in appropriate form. The reason can be extraction of charts can face issues of legal and copyright. Another reason might be missing relevant annotations on scrapping the charts from document. It can loss the necessary information. In the direction of converting extracted data into semi structured table Liang *et al.* [5] present a compositional semantic parsing approach for the ease of questions answering task .

Liu *et al.* [6] proposed a single deep neural classification framework that extract these textual and graphical information from bar and pie charts only with accuracy of 79.4 and 88.0%. while for other charts performance is very low. Savva *et al.* [7] proposed a Revision method to redesign charts for classification, visualization, and data extraction. This approach extracted low level images features using pixel's shape and color information for the purpose of image classification using SVMs. Bar charts text can be axis-tick, axis-title, legends, title and the data information of chart is available in the hight and width of bars. Various object detection methods like RCNN, Fast-RCNN, Faster-RCNN, YOLO, SSD and some specially designed methods have been proposed for the purpose of chart data detection and extraction in form of bounding box information.

Cliche *et al.* [8] presented a automatic data extraction from the scatter plot with 89% test accuracy. FigureQA is synthetic chart dataset proposed by Kahou *et al.* [9] of five class dataset. Most of these proposed dataset either have question answering pair annotations or have no annotations. No corpus is available for the chart to table and chart to latex pair annotations. When we see out of reasoning and QA task we found that image to latex script creation is also important field of research. Long Sort Term Memory LSTM [10] a variant of Recurrent Neural Network(RNN) [11] is widely used model to sequence to sequence problem for machine translation. A nice work in the filed of image to latex field is proposed by Deng *et al.* [12] to decompile an image to markup or latex markup. He introduces real world dataset of mathematical expressions with LaTex markup pair. He also proposed a synthetic dataset which include web pages and corresponding HTML snippet with higher performance with rendered images. Sutskever *et al.* [13] reported that his proposed method for sequence to sequence

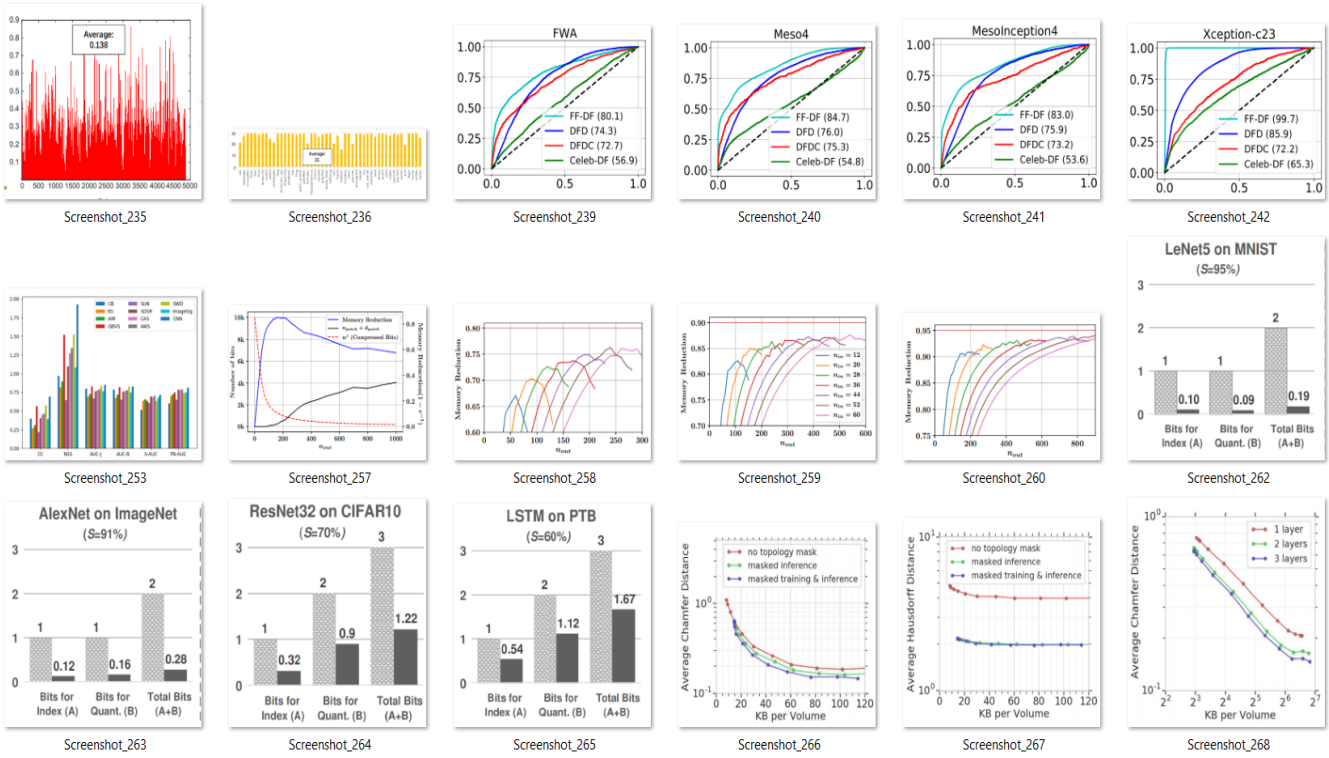

Figure 3. Our proposed dataset image gallery

A. *Challenges*

1) In [14], the DVQA dataset promoted the dynamic question encoding. However, this scheme shows more complexity to include pie-charts, plots and other visualizations since the dataset only contains the bar charts.

2) Although chart question answering pipeline modelled in [15], achieved reasonable accuracy, it failed to detect synonyms for visual features.

3) For Single Deep neural network developed in [6], augmenting the training data set by means of a more comprehensive simulation was an easy way to consistently improve the performance. However, it is more challenging to find a way to cope with small objects.

4) The method developed in [16], enhanced the performance of recalling of table selection model. But providing accurate decision on choice selection is still a tedious process.

5) In [17] , Deep learning model obtained significant improvements in training time when compared to Relation Networks. Moreover, on real life scientific figures QA tasks was not easy to accomplished.

III. The ChartReal Dataset

Lots of dataset presents in the field of Our newly created dataset is more realistic than existing datasets. We captured images from newly published paper in various fields and annotate that images. This dataset contains much more realistic and complex images and some of them are hard to annotate as well. This dataset contains almost all type of images from papers as they are not in same format.

Some of existing synthetic datasets like FigureQA, DVQA are more inclined toward reasoning over charts which do not have variable data labels, They do not have real valued data and not have annotated complex reasoning questions. PlotQA is more realistic dataset than FigureQA and DVQA but problem with PlotQA is following the same pattern of images. There is lack of variability of labels in this dataset too. Its annotation is also inclined to reasoning over charts. DocFigure [18] is dataset which contains real word document images of 28 categories mixed dataset and quality of these images are good but this dataset also not contain annotations figures and good for the purpose of classification of images.

As we are working on Sequence to Sequence machine translation, we require annotated table format of these images to generate the latex pair as group truth. No dataset

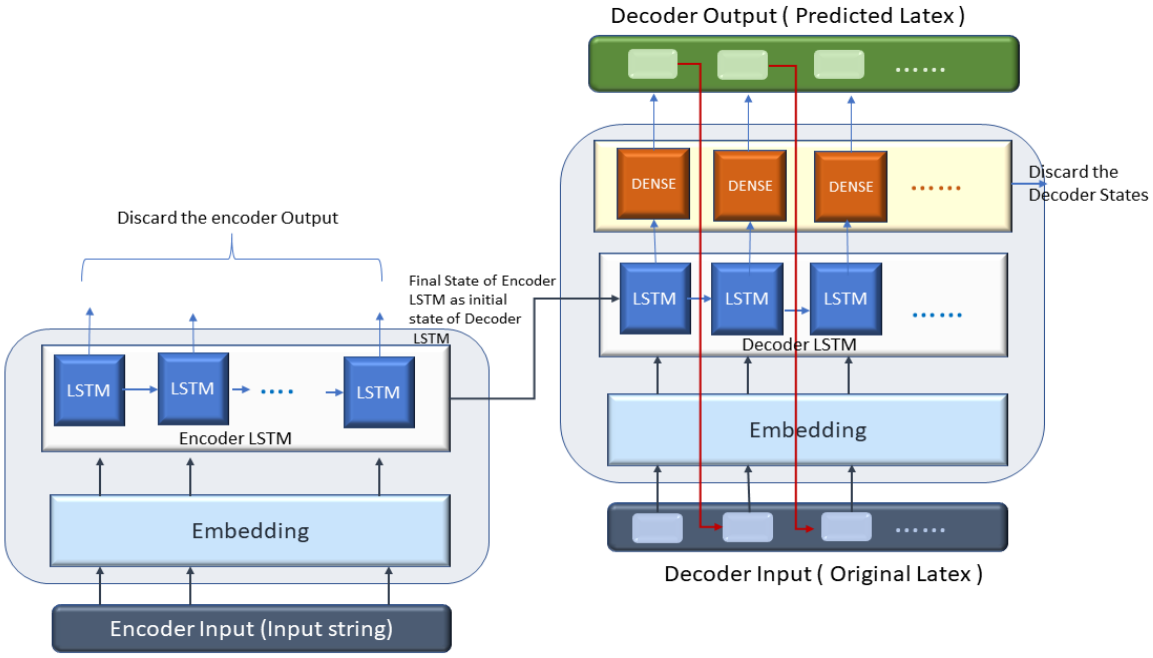

Figure 4. Methodology

contains charts annotations in table format. To provide the availability of table annotation of chart images, we have created and tested our model on 4 kind of dataset.

- First and Foremost, We captured and collected the images of size of 2k from the various standard research papers publicly available online. We annotated these charts into tables. In future we are targeting to make this dataset very large for further research.
- Secondly, We created synthetic dataset of 25k images to train our model. 2000 images are generated randomly for testing purpose. For creating synthetic images we used python script using matplotlib plotting libraries. This randomizes various variables like font, size, color, legends, x-tick, y-tick, number of images.
- We annotated the 40k images of bar chart out of 1 million dataset of PlotQA in table and latex pair out of which 40k images are bar charts images and ready to use as we are mainly focusing on bar chart in this paper.

IV. METHODOLOGY

A. Problem Statement

The Primary intention of this research will be to provide machine translation system for converting input string of bounding boxes and text labels from bar chart images to sequence of latex form.

B. *Feature extraction*

To detect text and position of bar, gray scaling and thresholding method of image segmentation in digital image processing provide values between 0(black) and 255(white). After binaries this value to black and white, Tesseract is used to extract the text from the image, along with their position, and appended into a single string, each reading is separated by two spaces.

The image is binarized and then inverted, following which, the opening (erosion followed by dilation) is performed on the image. This is done to get rid of all the noise and textual data, leaving us with only the bars of the chart. After the bars are obtained, the image is traversed along its axis to detect and calculate the height and position(w.r.t. to the leftmost corner) of the bar. All the readings obtained in this step are converted to a string and again append at the back of the string obtained from tesseract. The strings obtained in the two different steps are separated by four spaces.

This gives us a single string that contains the features of the graph and serves as an input to the seq2seq model in the next step.

C. *Data preprocessing*

The string that we obtain in the previous step is first passed through data cleaning and prepossessing where all the unwanted special characters, starting and ending white spaces were removed. The string that was obtained after

extracting the features from the graph image is passed on to the model as input after it is prepossessed. The model works on this string and predicts the output latex table script.

D. *Baseline 1: Pre-trained VGG16*

We used pre-trained VGG16 model by replacing top layer with task specific flatten layer as our first baseline model over 4000 synthetic training images and 1000 randomly generated testing images for each bar category to classify the charts into one of four categories of bar chart. Setting hyper parameters with batch size of 25, 50 epoch, SGD optimizer to train and validate the model. After classification of charts we used our written script for creating the table and latex script of table on extracted bar.

The problem we found with this method is we should have automatic model to convert images to latex script.

E. *Baseline 2: Sequence to Sequence model (Seq2Seq)*

This baseline makes use of the seq2seq model, an encoder-decoder based machine translation that takes an input text and converts it into another sequence of text. We used Seq2Seq Encoder-Decoder LSTM(Long Short Term Memory) based RNN Architecture. We used word level NMT (Neural Machine Translation). We made use of inbuilt embedding layers available in Keras API so that each word can be mapped into a fixed length vector.
Model Architecture generated using the utility of Keras:
The model was trained for 50 epochs and a batch size of 128 over a dataset generated using matplotlib in python over 25000 samples.

F. *Baseline 3: Image to Latex (Im2latex)*

V. PERFORMANCE EVALUATION

We ananlysed that our proposed solution is performing very well for the synthetic images dataset with the BLEU score of 0.48 as these images are following the similar pattern and are less variable. it contains four category of bar charts, horizontal bar charts, vertical bar charts, stacked horizontal and stacked vertical bar charts. But when testing the accuracy over another set of dataset we annotated, accuracy is not much satisfactory as it fells considerably. For the case of random generated dataset accuracy is low because of if various graphical effects to looks that images attractive enough. this happens. To improve the accuracy one should improve the Tesseract engine accuracy. because efficiency of engine is depending on the Tesseract engine. Another way to improve accuracy is using Recurrent Neural Network to extract text regions and using more accurate OCR. The loss is calculated as:

$$Error = \frac{|Value_{pred} - Value_{GT}|}{|Value_{GT}|}$$

VI. EXPERIMENTAL RESULTS

We used BLEU score [19] for the evaluation of quality of our translation. We analysed that the Model is predicting the correct table structure and labels in most cases on training data because as in the training data there were fewer labels as compared to 25k data. But it is not able to do numeric manipulations. But For New dataset with random string as labels, it fails to detect even the same table structure.

VII. CONCLUSION AND FUTURE WORK

We introduces the ChartReal dataset which is a collection of more realistic images than the previous synthetic dataset. We analysed the Sequence to Sequence models on our synthetic dataset and ChartReal dataset. This is challenging dataset and problem in document analysis for researchers. Our dataset is the initial base for future work. In future our object will be to get more accurate results and include all type of images from other books and news papers also. This is opening of new direction of research in this area.

REFERENCES

[1] Ray Smith and Google Inc. An overview of the tesseract ocr engine. In *Proc. 9th IEEE Intl. Conf. on Document Analysis and Recognition (ICDAR*, pages 629–633, 2007.

[2] Ahmad Pahlavan Tafti, Ahmadreza Baghaie, Mehdi Assefi, Hamid R. Arabnia, Zeyun Yu, and Peggy L. Peissig. OCR as a service: An experimental evaluation of google docs ocr, tesseract, ABBYY finereader, and transym.

[3] Xinxin Zhu, Weining Wang, Longteng Guo, and Jing Liu. Autocaption: Image captioning with neural architecture search, 2020.

[4] Harshit Rampal and Aman Mohanty. Efficient cnn-lstm based image captioning using neural network compression, 2020.

[5] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics.

[6] Xiaoyi Liu, Diego Klabjan, and Patrick NBless. Data extraction from charts via single deep neural network, 2019.

[7] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. Revision: Automated classification, analysis and redesign of chart images. In *ACM User Interface Software & Technology (UIST)*, 2011.

[8] Mathieu Cliche, David S. Rosenberg, Dhruv Madeka, and Connie Yee. Scatteract: Automated extraction of data from scatter plots. *CoRR*, abs/1704.06687, 2017.

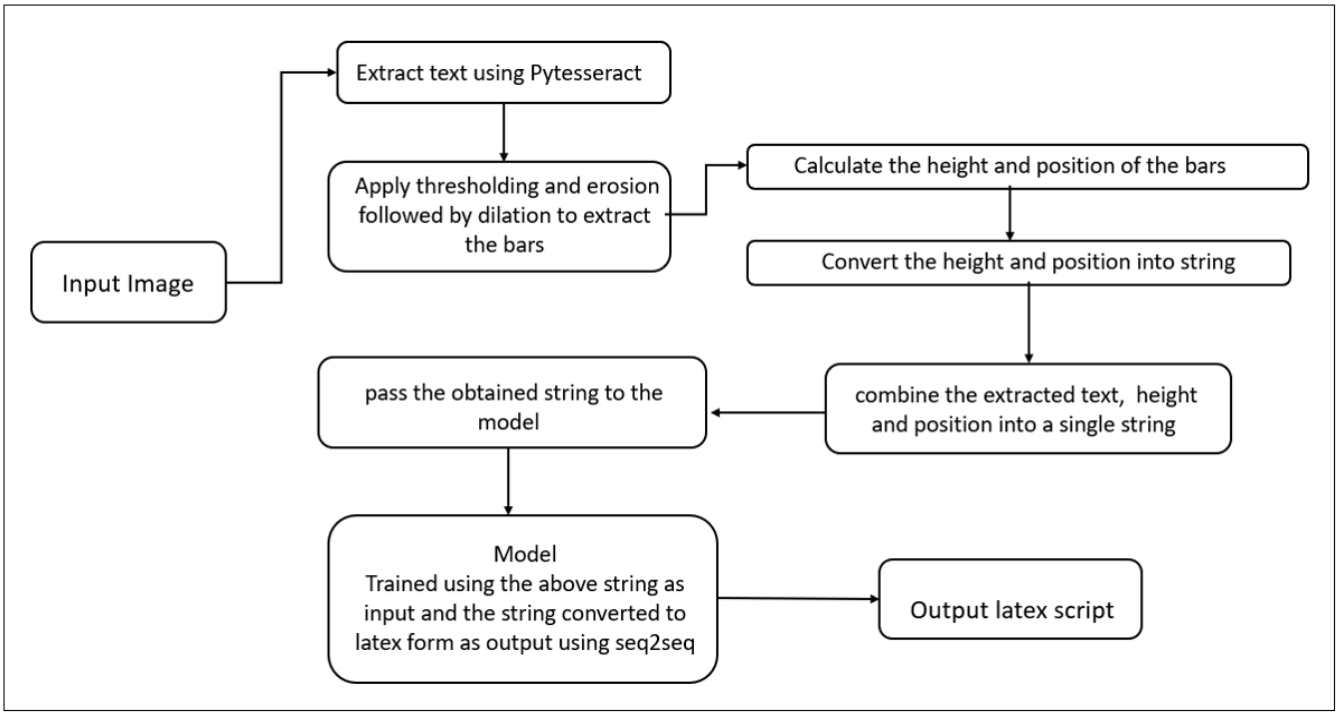

Figure 5. Baseline 2: Flow-Chart of proposed method showing that input image passing through various stages and converting into output latex script

[9] 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.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[11] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018.

[12] Yuntian Deng, Anssi Kanervisto, and Alexander M. Rush. What you get is what you see: A visual markup decompiler. abs/1609.04938, 2016.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc., 2014.

[14] Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. DVQA: understanding data visualizations via question answering. In *CVPR*, 2018.

[15] Hoque E. Kim, D.H. and pp. 1-13 April 2020 Agrawala, M. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Answering questions about charts and generating visual explanations.

[16] Hao Wang, Xiaodong Zhang, Shuming Ma, Xu Sun, Houfeng Wang, and Mengxiang Wang. A neural question answering model based on semi-structured tables. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1941–1951, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[17] Revanth Reddy, Rahul Ramesh, Ameet Deshpande, and Mitesh M. Khapra. Figurenet: A deep learning model for question-answering on scientific plots, 2019.

[18] K. V. Jobin, A. Mondal, and C. V. Jawahar. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79, 2019.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. 10 2002.