ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework

Junyu Luo*[1], Zekun Li*[2], Jinpeng Wang[3], and Chin-Yew Lin[3]

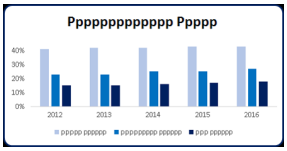
[1]Pennstate University, Pennsylvania, USA
[2]University of Southern California, Los Angeles, California, USA
[3]Microsoft Research, Beijing, China

Abstract

Chart images are commonly used for data visualization. Automatically reading the chart values is a key step for chart content understanding. Charts have a lot of variations in style (e.g. bar chart, line chart, pie chart and etc.), which makes pure rule-based data extraction methods difficult to handle. However, it is also improper to directly apply end-to-end deep learning solutions since these methods usually deal with specific types of charts. In this paper, we propose an unified method ChartOCR to extract data from various types of charts. We show that by combing deep framework and rule-based methods, we can achieve a satisfying generalization ability and obtain accurate and semantic-rich intermediate results. Our method extracts the key points that define the chart components. By adjusting the prior rules, the framework can be applied to different chart types. Experiments show that our method achieves state-of-the-art performance with fast processing speed on two public datasets. Besides, we also introduce and evaluate on a large dataset ExcelChart400K for training deep models on chart images. The code and the dataset are publicly available at `https://github.com/soap117/DeepRule`.

1. Introduction

Chart images can be easily found in news, web pages, company reports and scientific papers[24, 18, 10]. Automatic analysis of these data can bring us huge benefits, including scientific document processing, automatic risk assessment based on financial reports, and reading experience enhancement for visually impaired people. However, raw numerical tables are lost when charts are published as images. These underlying data of charts can be easily decoded by human, but not by machines [20]. Extracting the raw data table from chart images (see Figure 1 for an example) is the key step for understanding the chart content, which would lead to better analysis of related documents. Recent studies about question answering [12, 5, 13, 14] focusing on querying chart images would also benefit from it.



Input: Image of Chart

	Category0	Category1	Category2	Category3	Category4
Series0	40.58	41.55	41.56	42.34	42.35
Series1	22.66	22.67	24.66	24.68	26.43
Series2	15.02	15.05	15.88	16.85	17.66

Output: Data Table

Figure 1: Example of data extraction from a chart image

Some methods [1, 2, 20, 21] have been proposed for chart data extraction. These previous work heavily rely on manually crafted features. The diversity of chart designs and styles makes rule-based chart component extraction approaches difficult to scale. End-to-end solutions based on deep neural networks are also employed to tackle this problem because of their better accuracy [17, 6, 3], but these methods can not generalize well on all the chart types. For example, a framework designed for the pie chart cannot be applied to the line chart. Moreover, comparing to the heuristic rule-based methods, deep end-to-end approaches usually have no control of the intermediate results. Hence, a more general and flexible approach is desired to comprehend various chart images to further enhance the document analysis.

In this paper, we propose an approach that tackles the chart components detection problem with key point detection methods [15, 7, 16]. In this way, the chart data extraction can be simplified as a uniform task regardless of the styles of the chart images. Afterwards, an unified network is used for underlying data extraction. We design a deep hybrid framework that combines the advantages of

* Contribution during internship at Microsoft.