Looking Beyond Text:
Extracting Figures, Tables and Captions from Computer Science Papers

Christopher Clark and **Santosh Divvala**
The Allen Institute for Artificial Intelligence

http://pdffigures.allenai.org

Abstract

Identifying and extracting figures and tables along with their captions from scholarly articles is important both as a way of providing tools for article summarization, and as part of larger systems that seek to gain deeper, semantic understanding of these articles. While many "off-the-shelf" tools exist that can extract embedded images from these documents, e.g. PDFBox, Poppler, etc., these tools are unable to extract tables, captions, and figures composed of vector graphics. Our proposed approach analyzes the structure of individual pages of a document by detecting chunks of body text, and locates the areas wherein figures or tables could reside by reasoning about the empty regions within that text. This method can extract a wide variety of figures because it does not make strong assumptions about the format of the figures embedded in the document, as long as they can be differentiated from the main article's text. Our algorithm also demonstrates a caption-to-figure matching component that is effective even in cases where individual captions are adjacent to multiple figures. Our contribution also includes methods for leveraging particular consistency and formatting assumptions to identify titles, body text and captions within each article. We introduce a new dataset of 150 computer science papers along with ground truth labels for the locations of the figures, tables and captions within them. Our algorithm achieves 96% precision at 92% recall when tested against this dataset, surpassing previous state of the art. We release our dataset, code, and evaluation scripts on our project website for enabling future research.

1 Introduction

Mining knowledge from documents is a commonly pursued goal, but these efforts have primarily been focused on understanding text. Text mining is, however, an inherently limited approach since figures[1] often contain a crucial part of the information scholarly documents convey. Authors frequently use figures to compare their work to previous work, to convey the quantitative results of their experiments, or to provide visual aids to help readers understand their methods. For example, in the computer science literature it is often the case that authors report their final results in a table or line plot that compares their algorithm's performance against a baseline or previous work. Retrieving this crucial bit of information requires parsing the relevant figure, making purely text based approaches to understanding the content of such documents inevitably incomplete. Additionally, figures are powerful summarization tools. Readers can often get the gist of a paper by glancing through the figures which frequently contain both experimental results and explanatory diagrams of the paper's method. Detecting the associated caption of the figures along with their mentions throughout the rest of the text is an important component of this task. Captions and mentions help provide users with explanations of the graphics found and, for systems that seek to mine semantic knowledge from documents, captions and mentions can help upstream components determine what the extracted figures represent and how they should be interpretted.

Extracting figures requires addressing a few important challenges. First, our system should be ambivalent to the content of the figures in question, which means it should be able to extract figures even if they have heavy textual components, or are entirely composed of text. Therefore our algorithm needs to be highly effective at deciding when text is part of a figure or part of the body text. Second, we need to avoid extracting images that are not relevant (such as logos, mathematical symbols, or lines that are part of the paper's format), and to group individual graphical and textual elements together so they can all be associated as being part of the same figure. Finally, we seek to both identify captions and correctly assign figures and tables to the correct caption. Neither of these tasks is trivial due to the wide variety of ways captions can be formatted and the fact that individual captions can be adjacent to multiple figures making it ambiguous which figure they are referring to.

Our work demonstrates how these challenges can be overcome by taking advantage of prior knowledge of how scholarly documents are laid out. We introduce a number of novel techniques, including i) a method of removing false positives when detecting captions by leveraging a consistency assumption, ii) heuristics that are effective at separating body text and image text, and iii) the insight that a strong source of signal for detecting figures is the location of 'negative space' within the body text of a document. Unlike most previous work in this field, our system is equally effective at both table and figure detection.

Our system (pdffigures) takes as input scholarly documents in PDF form[2]. It outputs, for each figure, the bounding box of that figure's caption and a bounding box around the region of the page that contains all elements that caption refers to. Additionally we expect the correct identifier

[1] Throughout this paper we use the term 'figures' to refer to both tables, figures and their associated captions

[2] We assume the PDF format as it has become the de facto standard for scholarly articles