SciFig: A Scientific Figure Dataset for Figure Understanding

Anonymous Submission

Abstract

Most existing large-scale academic search engines are built to retrieve text-based information. However, there are no large-scale retrieval services for non-textual components such as scientific figures and tables. One challenge towards such services is scientific figure understanding that represents visual information by text. A key problem is a lack of datasets containing annotated scientific figures and tables, which can be used for classification, question-answering, and auto-captioning. Here, we design a pipeline that extracts figures and tables from scientific literature and a deep-learning-based framework that classifies scientific figures using visual features. Using this pipeline, we develop the first large-scale annotated corpus, SCIFIG consisting of more than 264k scientific figures extracted from $\approx$ 56k research papers in the ACL Anthology. We make available the SCIFIG-PILOT dataset that contains 1671 manually labeled scientific figures belonging to 19 different categories. The dataset is accessible at https://bit.ly/3m4u0eq under a CC BY-NC license.

1 Introduction

Figures are ubiquitous in scientific papers to illustrate experimental and analytical results. We refer to these figures as *scientific figures* to distinguish them from natural images, which usually contain richer colors and gradients. Scientific figures provide a compact way to present numerical and categorical data, which often enable researchers to draw more intuitive insights and conclusions. Automatic understanding of scientific figures can assist in developing retrieval systems that discover from hundreds of millions of papers that are readily available on the Web (Khabsa and Giles, 2014). The state-of-the-art machine learning models can read captions and parse shallow semantics for certain types of scientific figures. However, the task of building a general and robust system that can reliably represent and interpret visual information and connect it with text content remains a challenge. One key step to facilitate advancing figure understanding is to build datasets containing diverse collections of scientific figures and their textual descriptions.

Here, we propose a pipeline to build a categorized and contextualized scientific figure dataset. Applying the pipeline on 55,760 papers in the ACL Anthology (downloaded from https://aclanthology.org/ in mid-2021) we built two datasets: SCIFIG and SCIFIG-PILOT. SCIFIG consists of 263,952 scientific figures, their captions, inline references, and metadata. SCIFIG-PILOT is a subset of SCIFIG, consisting of 1671 scientific figures. It was manually classified into 19 categories. The SCIFIG-PILOT dataset can be used as a benchmark for scientific figure classification. The pipeline is open source and configurable, enabling others to expand the datasets by extracting and annotating figures from other scholarly datasets with pre-defined or new labels.

2 Related Work

2.1 Scientific Figures Extraction

Automatically extracting figures from scientific papers is important because many downstream tasks rely on large numbers of accurately extracted figures. Wu et al. (2015) introduced a multi-entity extraction system called PDFMEF, incorporating a scientific figure extraction module. Shared tasks such as ImageCLEF (Müller et al., 2015) drew attention to compound figure detection (Yu et al., 2017) and separation (Tsutsui and Crandall, 2017). Clark and Divvala (2015) proposed a framework called PDFFIGURES that extracted figures and their captions in research papers. The authors extended their work and built a more robust framework called PDFFIGURES2 (Clark and Divvala, 2016). DEEPFIGURES was later proposed to overcome the limitations of the above frameworks by incorporating deep neural networks, i.e.,ResNet-101 (Siegel et al., 2018a).

2.2 Scientific Figure Classification

Scientific figure classification (Savva et al., 2011; Choudhury and Giles, 2015) helps machine understanding of figures. Early work used a visual bag-of-words representation with a support vector machine (SVM) classifier (Savva et al., 2011). Hough transforms recognized bar charts in document images (Zhou and Tan, 2000b,a). Prasad et al. considered Scale Invariant Feature Transform (SIFT) (Lowe, 2004) and Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) features to recognize five different types of charts (Prasad et al., 2007). Handcrafted features were used to classify charts in scientific documents into various types, e.g., Zhou and Tan (2000b); Siegel et al. (2016); Vitaladevuni et al. (2007). However, handcrafted features usually did not generalize well. As such a convolutional neural network (CNN)-based model was proposed (Kavasidis et al., 2018) which identified the locations of tables, bar charts, and pie charts in research papers. Another that combined CNN and the deep belief networks showed improved performance compared with feature-based classifiers (Tang et al., 2016).

Table 1: Datasets for scientific figure classification.

Dataset	Labels	#Figures	Image Source
Deepchart	5	5000	Web Image
Figureseer	5	30600[1]	Web Image
Prasad et al.	5	653	Web Image
DocFigure	28	33000[2]	Scientific Papers
Revision	10	2000	Web Image
FigureQA[3]	5	100000	Synthetic figures
SciFig-pilot	**19**	**1671**	**Scientific Papers**
SciFig[4]	-	**263952**	**Scientific Papers**

[1] Only 1000 images are public.
[2] Not publicly available.
[3] Scientific-style synthesized data.
[4] SciFig does not contain human-assigned labels.

2.3 Figure classification Datasets

Existing datasets for figure classification include DocFigure (Jobin et al., 2019), FigureSeer (Siegel et al., 2016), Revision (Savva et al., 2011), and datasets presented by Karthikeyani and Nagarajan (2012) and Vitaladevuni et al. (2007). Most datasets were collected from the Web except for DocFigure, which was created by extracting figures from scientific documents. FigureSeer and DocFigure each contain more than 30k images. The sizes of other datasets are relatively small. Only a small subset ($\approx$ 1000) of the FigureSeer dataset was labeled. Most datasets have no more than 10 labels except for DocFigure, which has 28 labels. Table 1 summarizes existing datasets that may be used for scientific figure classification.

FigureQA is a dataset consisting of over one million question-answer pairs grounded in over 100,000 synthesized scientific images (Kahou et al., 2018) with five styles. Our dataset is different from FigureQA because the figures were directly extracted from research papers.

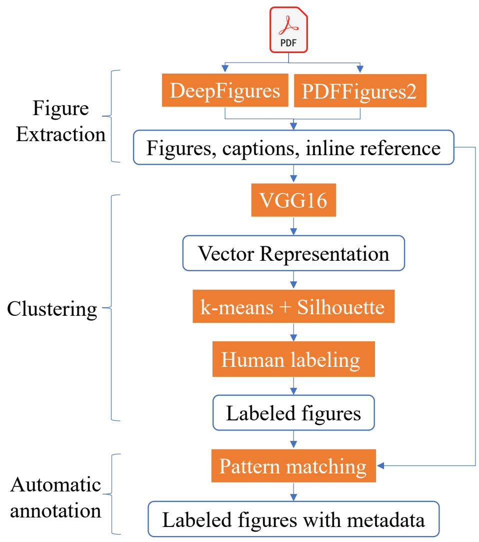

Figure 1: Overview of the data generation pipeline.

3 Data Generation Methods

The ACL Anthology corpus is a sizable, well-maintained PDF corpus with clean metadata covering papers in computational linguistics with freely available full-text. Previous work on figure classification used a set of pre-defined categories, e.g., Kahou et al. (2018), which may not cover all types of figures. We use an unsupervised method to determine figure categories. After the category label is assigned, each figure is automatically annotated with metadata, captions, and inline references. The pipeline includes 3 steps: figure extraction, clustering, and automatic annotation. An overview of the data generation pipeline is illustrated in Figure 1.

3.1 Figure Extraction

We extracted figures using PDFFIGURES2 and DEEPFIGURES. PDFFIGURES2 (Clark and Divvala, 2016) first identifies captions and the body text inside a document, because these elements can often be identified accurately in scientific articles. Areas containing figures can then be located by identifying rectangular regions adjacent to captions and not overlapped with the body text.

DeepFigures (Siegel et al., 2018b) uses the distant supervised learning method to induce labels of figures from a large collection of scientific documents in LaTeX and XML format. The model is based on TensorBox, applying the Overfeat detection architecture to image embeddings generated using ResNet-101 (Siegel et al., 2018a). We utilized the publicly available model weights[1] trained on 4M induced figures and 1M induced tables for extraction. The model outputs the bounding boxes of figures and tables. Here, unless otherwise, we refer to figures and tables as "figures".

Using DeepFigures and PDFFigures2, we successfully extracted 249,669 figures and 254,906 figures from 55,760 papers, respectively. Each process extracts figures following the steps below. The system extracts figures at a rate of 200 papers per minute on a Linux server with 24 cores.

1. Retrieve a paper identifier from the job queue.
2. Pull the paper from the file system.
3. Extract figures and captions from the paper.
4. Crop the figures out of the rendered PDFs using detected bounding boxes.
5. Save cropped figures into PNG format and the metadata in JSON format.

3.2 Clustering Methods

Now we use an unsupervised method to classify extracted figures. We extract visual features using VGG16 (Simonyan and Zisserman, 2015), pre-trained on the ImageNet dataset (Deng et al., 2009). VGG16 contains a series of convolutional layers followed by max-pooling layers and a set of 3 fully connected dense layers. VGG16 has been used in document representation learning and pattern analysis, e.g., (Simonyan and Zisserman, 2014).

All input figures are scaled to a dimension of 224×224 to be compatible with the input requirement of VGG16. The features were extracted from the second last hidden (dense) layer, consisting of 4096 features. Principal Component Analysis was adopted to reduce the dimension to 1000.

Next, we cluster figures represented by the 1000-dimension vectors. We compare two heuristic methods to determine the optimal number of clusters, including the Elbow Method (Thorndike, 1953) and the Silhouette Analysis (Rousseeuw, 1987). To use the method, one needs to examine the *explained variation*, a measure that quantifies the difference between the between-group variance to the total variance, as a function of the number of clusters. The pivot point (elbow) of the curve determines the number of clusters to use.

Silhouette Analysis determines the number of clusters by measuring the distance between clusters. The Silhouette plot displays how close each point in one cluster is to points in the neighboring clusters, allowing us to visually assess the cluster number. This measure has a range of $[-1, 1]$. Silhouette Analysis takes into account more factors, e.g., variance, skewness, and high-low differences, and is usually considered a better method.

[1]https://github.com/allenai/deepfigures-open

3.3 Automatic Annotation

This automatically associates figures to metadata, including captions, inline reference, figure type, figure boundary coordinates, caption boundary coordinates, and image text (text appearing on figures, only available for results from PDFFigures2). The figure type is determined in the clustering step above. The inline reference is obtained using GROBID (see below). The other metadata was available in the output of the figure extractor. PDFFigures2 and DeepFigures extract the same metadata fields except for "image text" and "regionless captions" (captions for which no figure regions were found), which are only available for results of PDFFigures2.

An inline reference is a text span that contains a citation to a cross-reference, such as a figure or a table. Inline references can be useful to understand the relationship between text and the entities it refers to. After processing a paper, GROBID outputs a TEI file (a type of XML file), containing marked-up full-text and references. We locate inline references of a particular figure using its label (e.g., "Figure 1") and extract the sentence containing the label. A regular expression was used to match figure labels.

4 Results

4.1 Figure Extraction

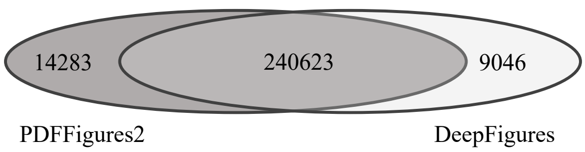

Figure 2: Numbers of extracted figures.

We use both PDFFigures2 and DeepFigures to extract figures. The numbers of extracted figures

Table 2: Figure class distribution in the SCIFIG dataset.

Class	%	Class	%
Trees	13	Graphs	6
Natural Images	8	Tables	6
Confusion Matrix	7	Screenshots	6
Pie Charts	6	Scatter Plot	4
Bar Charts	6	Maps	3
NLP text/grammar	6	Boxplots	2
Architecture Diagram	6	Venn Diagram	1
Algorithm	6	Word Cloud	1
Neural Networks	6	Pareto	1
Line Graph	6		

by these two packages are shown in Figure 2. The diagram indicated that there is a significant overlap between figures extracted by both software packages. However, each package extracted ($\approx 5\%$) figures that were not extracted by the other package. By inspecting a random sample of figures extracted by both software packages, we found that DEEPFIGURES tended to miss cases in which two figures were vertically adjacent to each other. We took the union of all figures extracted by both software packages to build the SCIFIG dataset, which contains a total of 263,952 figures. All figures extracted are converted to 100 DPI using standard OpenCV libraries. The total size of the data is $\sim$ 25GB before compression. Inline References were extracted using GROBID wrapped by PDFMEF. About 78% of figures have inline references.

4.2 Determining the Cluster Number

The extraction contains $\sim$ 150k tables and 110k figures. The figures were clustered using the k-means algorithm. We increased k from 2 to 20 with an increment of 1 to determine the number of clusters. The results were analyzed using the Elbow Method and Silhouette Analysis. No evident arm was observed in the Elbow Method. The Silhouette diagram exhibits an evident turning point at $k = 15$, where the score reaches the maximum. Therefore, we group the figures into 15 clusters. To validate the clustering results, 100 figures randomly sampled are manually inspected from each cluster. We identified three additional types of figures, namely *word cloud*, *pareto*, and *venn diagram*. The SCIFIG-PILOT dataset was built using these manually inspected figures. For completeness, we add 100 randomly selected tables. Now the SCIFIG-PILOT dataset contains a total of 1671 figures and tables in 19 classes. The distribution of all classes is shown in Table 2. Examples of figures are shown in Figure 3.

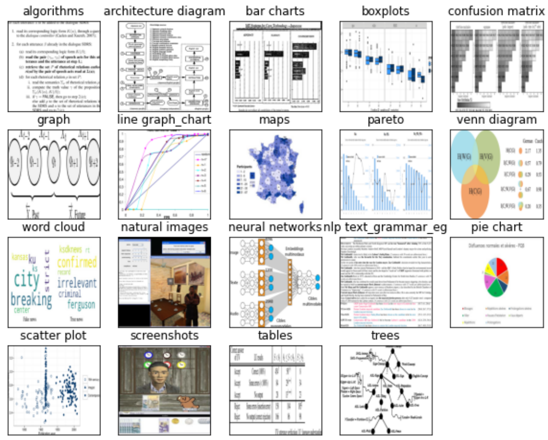

Figure 3: Example figures in the SCIFIG-PILOT dataset.

5 Figure Classification

Based on the SCIFIG-PILOT dataset, we train a supervised classifier. The dataset was split into a training and testing set with an 8:2 ratio. Two deep learning models were investigated. The first model is a 3-Layer CNN, trained with a categorical cross-entropy loss function and the Adam optimizer. The model contains three typical convolutional layers, each followed by a max-pooling and a drop-out layer, and three fully-connected layers. The dimensions are reduced from 32×32 to 16×16 to 8×8. The last fully connected layer classifies the encoded vector into 19 classes. The classifier achieves an accuracy of 59%. The second model was trained based on the VGG16 model (Simonyan and Zisserman, 2014) except that the three fully-connected layers at the top of the original network were replaced by a long short-term memory layer, followed by several dense layers for classification. This model achieves an accuracy of $\sim$ 79%, 20% higher than the 3-Layer CNN model.

6 Conclusion

We designed a pipeline that builds a corpus of classified scientific figures and applied it to ACL Anthology papers leveraging state-of-the-art figure extraction frameworks. This corpus, SCIFIG, consists of $\approx$ 250k scientific figures and tables, and SCIFIG-PILOT, a subset of SCIFIG, consisting of 1671 scientific figures with 19 manually verified labels. One limitation of our pipeline is the determination of the number of clusters required visual inspection. Future work could be using density-based methods, e.g., Xuanzuo et al. (2017), to fully automate the clustering module.

References

Sagnik Ray Choudhury and C. Lee Giles. 2015. An architecture for information extraction from figures in digital libraries. *Proceedings of the 24th International Conference on World Wide Web*.

Christopher Clark and S. Divvala. 2015. Looking beyond text: Extracting figures, tables and captions from computer science papers. In *AAAI Workshop: Scholarly Big Data*.

Christopher Clark and S. Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 143–152.

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning.

V. Karthikeyani and S. Nagarajan. 2012. Machine learning classification algorithms to recognize chart types in portable document format (pdf) files. *International Journal of Computer Applications*, 39:1–5.

I. Kavasidis, S. Palazzo, C. Spampinato, C. Pino, D. Giordano, D. Giuffrida, and P. Messina. 2018. A saliency-based convolutional neural network for table and chart detection in digitized documents.

Madian Khabsa and C. Lee Giles. 2014. The number of scholarly documents on the public web. *PLoS ONE*, 9(5):e93949.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Henning Müller, Alba García Seco de Herrera, and Stefano Bromuri. 2015. Overview of the imageclef 2015 medical classification task. *Workshop proceedings of the 6th International Conference and Labs of the Evaluation Forum Association 2015 : Experimental IR Meets Multilinguality, Multimodality, and Interaction*, (CONFERENCE):13 p.

V. Shiv Naga Prasad, Behjat Siddiquie, Jennifer Golbeck, and Larry S. Davis. 2007. Classifying computer generated charts. In *2007 International Workshop on Content-Based Multimedia Indexing*, pages 85–92.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

M. Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and J. Heer. 2011. Revision: automated classification, analysis and redesign of chart images. *Proceedings of the 24th annual ACM symposium on User interface software and technology*.

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

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018a. Extracting scientific figures with distantly supervised neural networks. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*.

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018b. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, pages 223–232. ACM.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Binbin Tang, Xiao Liu, Jie Lei, Mingli Song, Dapeng Tao, Shuifa Sun, and Fangmin Dong. 2016. Deepchart: Combining deep convolutional networks and deep belief networks in chart classification. *Signal Processing*, 124:156–161. Big Data Meets Multimedia Analytics.

Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18(4):267–276.

Satoshi Tsutsui and David J. Crandall. 2017. A data driven approach for compound figure separation using convolutional neural networks. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:533–540.

S. Vitaladevuni, Behjat Siddiquie, J. Golbeck, and L. Davis. 2007. Classifying computer generated charts. *2007 International Workshop on Content-Based Multimedia Indexing*, pages 85–92.

Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C. Lee Giles. 2015. PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search. New York, NY, USA. Association for Computing Machinery.

Ye Xuanzuo, Li Dinghao, and He Xiongxiong. 2017. An algorithm for automatic recognition of cluster centers based on local density clustering. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 1347–1351.

Yuhai Yu, Hongfei Lin, Jiana Meng, Xiaocong Wei, and Zhehuan Zhao. 2017. Assembling deep neural networks for medical compound figure detection. *Inf.*, 8:48.

Y. Zhou and C. Tan. 2000a. Bar charts recognition using hough based syntactic segmentation. In *Diagrams*.

Y. Zhou and C. Tan. 2000b. Hough technique for bar charts detection and recognition in document images. *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, 2:605–608 vol.2.