International Journal of Engineering Sciences Paradigms and Researches (Vol. 39, Issue 01) and (Publishing Month: Feb 2017)

Imagery Dictionary using CNN Approach to Assist Visually Challenged People

Gati Krushna Nayak[1], Narendra Kumar Rout[2] and Nibedita Sahoo[3]
[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar
[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Publishing Date: February 27, 2017

Abstract

Our work aims to enable the visual information in a document to be accessible by the visually impaired or the blind people. The blind people prefer arts over science subjects in higher education because conveying equations and algorithms to them is seen as difficult. They should not be deprived of acquiring knowledge due to physical disabilities. In this work, we identify different types of images in an algorithm textbook as graphs, algorithm images, equations, and network flow diagrams with 98% accuracy. The Convolution Neural Networks used can predict new data instances with 99% accuracy except for recurrence tree images. We want to extend our work to extract the textual information in these images and produce image descriptions as alt text. This can produce a better speech output to assist the blind people.

Keywords: *Assistive Study, Visually Challenged, CNN, Visually Challenged, Imagery Dictionary.*

I. INTRODUCTION

The blind people get information through other sensory perceptions like touch, smell and hearing. They read and perceive information through the printed documents in the form of Braille, and they can access the digital contents through screen reader using text to speech converters. Reading and interpreting the textual content have many solutions but how to transfer the graphical content in the documents is a real challenge. Most of the documents have graphical data as tables, charts, bar graphs or pie charts or scatter plots. Interpreting graphical information by the visually impaired is very difficult and nearly impossible for the blind. The person who has lost his sight in a later stage of life can interpret it more easily compared to a congenital blind.

We present an approach for classifying pictures in the computer science algorithm text book [1] into different categories. We made eight classes including bar charts, pie charts, line graphs, algorithm images, recurrence trees, equations etc. This classification would be useful for further extraction of information from figures and also would like to extend this as alt text or alternative text. Alt text is the image description inserted in the HTML(Hyper Text Markup Language) and can be read out by screen readers used by the visually impaired.

This paper is organised such that motivation and detailed survey of the literature studies is given in section II, the method used in section III, results in section IV and section V includes the summary and future work followed by references.

II. Literature Survey

Our paper aims to classify the different types of 'informative figures' in a document and parse it to produce alt text from the information obtained. This alt text is readable by a screen reading software and thereby reaching the blind people. We would also like to classify all the non-text information in a document, and identify the embedded data and thus help the blind in education.

Most of the research works converts the graphical data to a format which can be easily perceived by the blind mostly as haptics or audio format.

- Haptic output- In haptics, perception by the blind is through the vibration or force feedback from the devices. Yu Brewster et al. (2003) in [2] evaluated the interpretation of graphs by the blind people. They took the charts in the digital documents and converted to the haptic model using a SenSable Phantom, and a Logitech WingMan force feed-back Mouse. The user can control the pointer on the screen and feel the force generated by these devices. Moustakas et al. [3] converts 2D map image or 3D video of maps to haptics and audio output and in this, they solve the challenge of the cross-modal transformation of visual data into haptic.
- Audio output- The aural output is mostly as speech or non-speech output like music. The audio representation assists them in understanding the contents. [2] uses Microsoft Speech SDK7.0 for speech output and also use the musical note in which high pitch amount to a higher data value and a low pitch mapped to lower data value. In [3] they use TTS.

They were able to do so successfully for lines and simple bar charts. The blind were asked to map the graphs on the screen to a paper as they perceive through audio or tactile devices to evaluate if the mental sketches match with the chart. As the number of lines increases in the graph, it was difficult to form an overview due to the narrow bandwidth on Phantom and imparting the exact data values to the blind was difficult. But integrating auditory feedback was better than haptics alone. The bar graph is also implemented in describing the overall trend of the data for highest and lowest bars, but found it challenging to find height between closely related bars.

TABLE I EXISTING WORKS IN ANALYSING GRAPHS IN DOCUMENTS

Author	Year	Key Approach	Input	Output
Kahaou et al [8]	2018	Created five classes of graphs and their corresponding question answer pairs from the source data	Graph synthesised using Bokeh [20]	Binary output related to the question asked
Seigel et al [7]	2016	Text localisation and identification by Microsoft's OCR	The graphs extracted from research paper Classification based on pixel based features and CNN based features.	Classify the graphs and analyse the con-tent
Abhijit et al [9]	2018	Created Dataset using Matplotlib Charts classified by CNN architecture Text boxes identified by bounding boxes OCR using Google Tesseract Performance of object detection by mAp estimation	Five types of Chart images in PNG format	Textual description of the charts created automatically to produce alt text.
Yui Kita et al. [10]	2017	Random Forest and SVM	Research Paper	Predict the importanceof figures in scholarly documents
Silvia et al. [11]	2017	Accessiblity of graphs in HTML based articles. Automatically generating graphs from tabular data	The csv documents in RASH(Research Article in Simplified HTML format)	Read it out by Voice Over in MacOS and JAWS, NVDA in Windows
Stephanie et al [12]	2007	Bar charts made accessible by making a browser extension. Only for electronics images produced within a specified format. Image processing for detecting BAR graphs information about Bar graph is augmented in xml and Bayesian Inference for recognising intendation of the graphic.	Bar graphs in the web pages	Summarise the graphics as text and read it to user by JAWS
Wai Yu et al [13]	2003	Virtual graphs rendered by computer and the blind percieve through hardware devices.	Graphics on the screen-line and Bar graph	Data read out as speech or Non speech as music with varying pitch.
Clarks et al [14]	2015	Identify titles, body text and captions in a pdf. Made a dataset composing of 150 research papers in computer science.	Scholarly documents in PDF format	Finds bounding box of the around the caption and also find bounding box in a region related to the caption.
Savva et al. [16]	2011	Classification of chart images. Extract underlying data from the charts and redesign it to improve the perception. Employ manual approach in annotating the text regions	Input is bitmap images of charts.	View alternate chart designs. Successful in bar and pie charts.
Clark et al. [17]	2016	An algorithm called PDFFigures2.0 for mining diagrams from the research papers. Introduces a dataset with 150 computer science papers with ground truth. Region identification divides the paper into caption, region with text, figures and text associated with images.	Computer science research papers	Figures, tables and captions extracted from the paper along with their page number
Al-Zaidy et al. [18]	2015	Graphical extraction module extracts the bar using connected component and lab color spaces. Text and numerical values is identified by tesseract OCR. Parsing of legends is a possible future extension.	2D bar and pie charts	Extract the bar chart from pdf, web images

The blind are elucidating the graphical data with the help of an assistant or a volunteer. If they use screen readers, it can read out the text along the axes and screen reader have difficulty in reading content in the data in tabular form. The charts in the web pages are made accessible to the visually impaired by Alison et al. (2018) in [4]. The data in the graph is converted to image format or pdf with the help of Highcharts, that is a third-party software library that takes in data as SVG or JSON and turns to image or pdf. The screen reader could read out the descriptions of the figure which is added by the chart creator and also automated illustrations of the type of charts, and details about the axis the contents of the web are accessed with the help of screen readers. JAWS is a screen reading software for Windows developed by Microsoft and

Voice-over in MacOS. The caret browsing help to choose the various text on the web with the keyboard.

The web accessibility is made successful with the 'alt text' or alternative text. Alt text is used in HTML code to describe the appearance of an image in a web page to people who can't see. [5] discusses that the majority of web pages are without alt text. The various issues faced by the blind user accessing the web content appears in [6] and most of the sites do not provide alt text for the pictures and charts. Elzer et al. (2007) [5] they made a browser extension in the internet explorer to identify the bar chart image on the web page. Detects it as a bar graph if there are at least two rectangles with the same row or column. For that is not a bar graph, it reads as not a bar graph by JAWS, and this works for electronic documents with a predefined set of fonts and no overlapping characters.

Prasad et al. in [15] classify computer-generated charts using the spatial and shape features fed into an SVM. In this, each pair of an image is checked for similarity by pyramid matching algorithm. Savva et al.(2011) in [16] extracted data from the graphs and redesigned it to improve the perception. They also employed a manual approach to annotating the text region and looks forward to a full automated annotation in future.

Most of the analysis of results in research papers is as figures. The results as graphs in research papers are extracted automatically, classified and then analysed in FigureSeer [7] by Seigel et al. (2016). The graph plots from the document are taken and the axes labels, position and scales are analysed. Microsoft OCR does text localisation and identification. The legend label is a pair (legends, symbols) where former is a variable and symbols give its appearance. Finding label is a text classification problem dealt with random forest where the input is six-dimensional feature corresponding to each text box symbols found by using two rectangular boxes having a score assigned with non-background pixel density. Pixelbased features followed by SVM analyses the plot data, and this showed a lesser accuracy than CNN based classification. Done on Titan X GPU using Caffe and made an annotated figure parsing dataset called FigureSeer and find applications for query answering related to the graphical plots in research papers. In research papers comparison with other related works are presented mostly as tables and graphs. The pictures are extracted from scholarly documents but identifying the text if it's the part of the figure or not, is crucial.

TABLE II: EXISTING DATASETS FOR DIFFERENT TYPES OF GRAPHS AND RELATED WORKS

Dataset	Year	Details	Approach
FigureQA [8]	2018	Synthetic graphical data belonging to five classes and their question answer pair	Four models used a)LSTM: for text alone b)CNN + LSTM: visual and text c)A baseline VGG16 d)Relation network
FigureSeer [7]	2016	Original images extracted from research papers. Localising figures by bounding boxes.	The figure parsing uses text localisation and identification by Microsoft OCR followed by classification as a)Pixel based approach, pair wise petential to know the smooth transitions among adjacent pixels b)CNN based similarity features and evaluated AlexNet and ResNet.
CS150 [14]	2015	Consists of 150 computer science papers along with ground truth for locations of figures and tables	Academic papers follow a strict convention. a)The algorithm does caption identification but restricted to first word of caption. b)Region identification in which the whole pdf is divided by bounding boxes. c)Figure assigned to its corresponding caption by pixel based region scoring.

Clarks et al. (2015) in [14] for the documents in PDF form, finds bounding box of the around the caption and also find bounding box in a region related to the caption. The scholarly do the comparison with related work or results as tables or figures. They resolve the challenge of extracting figures with high textual content and also avoid logos and symbols that are part of the structured scholarly documents. They were successful in identifying captions, images along with text and tables. For the input PDF, a bounding box around the caption of the figure is formed and they made a dataset with 150 computer science related papers. Since charts are available as images and the basic data is not known to the users. Manolis et al. in [16] classify the chart images into different classes except for the one with 3D effect and shading, using SVM and finds the mark and corresponding data and redesigns the chart for better visual understanding. Table I gives a detailed summary of the existing works in analysing graphs in documents.

A. Datasets available

Any significant advancements in the research requires public datasets for training. The table II gives an insight to available datasets on graph images and summarises the features used by them.

III. METHOD

The dataset includes of 200 horizontal bar charts, 200 vertical bar charts, 200 line graphs, 200 pie charts and 200 scatterplots from the figureqa [8]. The data also consists of images collected from the CLRS book and augmenting it with pictures from the internet. 200 algorithm images, 200 equations as images, 200 network flow diagrams. The input image is reshaped to 200x200 size and converted to grayscale as this deals with document images.

A. CNN Architecture

We use deep learning to classify the images using Convolution Neural Networks or CNN. ConvNet is a kind of neural network which can recognise the visual patterns in a picture just like a human brain. ConvNet differs from another neutral network as it should have at least one convolution layer [19].CNN make use of locality of pixel dependencies. The convolution operation extracts the low-level features like edges initially, and more abstraction happens through different convolution layers towards recognising the image. We choose hyper parameter tuning of the CNN. Keep changing the parameters until no further decrease in the loss.

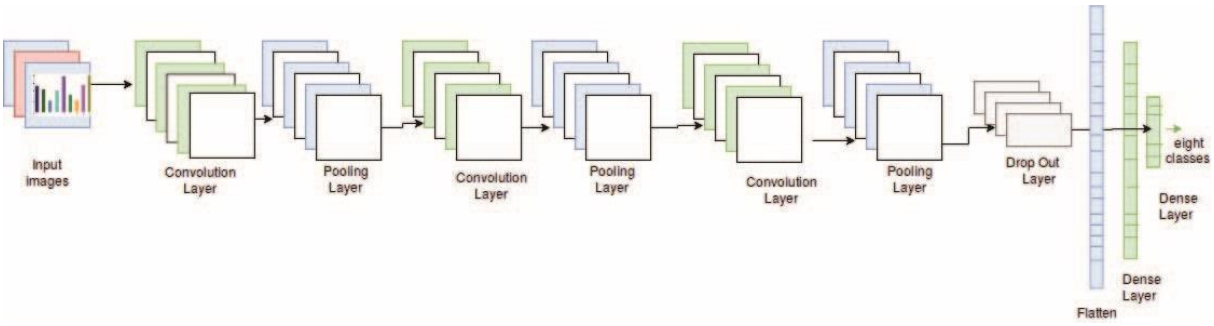


Fig. 1. CNN Architecture

B. Tuning of CNN

The architecture of our model has nine layers- 3 convolution layers, three max-pooling layers, on drop out and two fully connected layers. Fig.1 shows the architecture of the model. Fine tuning includes setting of the hyper parameters of the network, and is difficult as training process is slow and this includes the choice of number of layers , activation function, learning rate, drop out, number of epochs and batchsize.

We use 32 filters of size 3x3 which scans over the image to produce 32 sets of activation map. We choose 1x1 or 3x3 filters usually, 3x3 is a good choice since the stride is of one pixel. The filter shifts by one-pixel position when stride is chosen as one. As the filter size increases, it extracts more general feature. When the kernel size is 1x1, it is similar to treating all the pixels as an essential feature. So choosing an optimal kernel size is vital.

1) Pooling Layer: Pooling selects the maximum or minimum pixel value in the defined 3x3 neighbourhood using sliding window approach. Each convolution layer is followed by a max-pooling layer for subsampling and pooling layer operates on each feature map independently.

2) Activation Function: The activation function introduces non-linearity to the whole system. Out of the different types of activations possible, our system uses ReLu and Softmax. • ReLu : Rectified Linear Unit follows every convolution layer. Mathematically, y = max (0, z) .

$$R(z) = \begin{cases} z & ; \quad z > 0 \\ 0 & ; \quad z \leq 0 \end{cases}$$

It makes all the negative values to zero. • SoftMax: It is used at the final dense layer. It finds the probability of *n* different target classes over all the possible classes.

3) Dropout Layer: Dropout is a regularization method for reducing over fitting in the fully connected or the dense layers. If there are *n* units in the hidden layer and *p* is the probability of keeping the neuron, then *pn* units will be present in the layer after dropout [2].

A flattening follows the drop out layer. The output of the hidden layers(that is convolution and pooling) is converted into a 1D feature vector before feeding into the fully connected layer, as the dense layer is just a Neural Network.

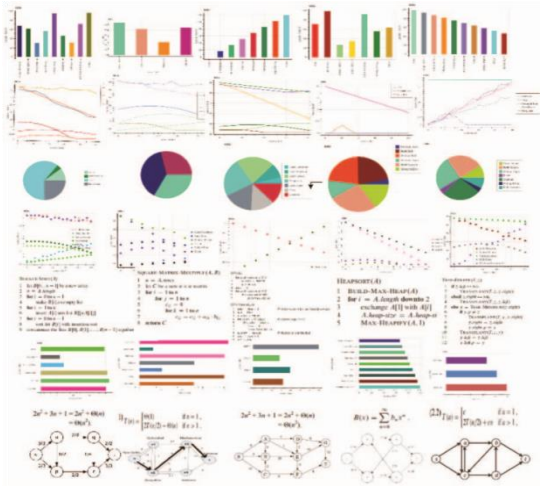


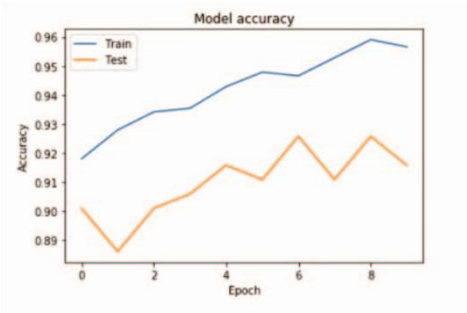
Fig. 2. Sample images from the dataset. Each row corresponding to different

We classified the diagrams using CNN on python3 with the tensorflow backend, and our model is capable of categorising with 98% accuracy. The summary of our model is given in Fig.3 detailing the feature maps after each layer. The model accuracy on training and validation dataset is clear from the learning curve in fig.4. The model accuracy and loss is comparable in both training and validation phase, except for a hike in loss or depression in performance at the hundredth epoch, and thus we made training restricted to earlier epochs. We made predictions on new data instances by taking fifteen images form the internet, corresponding to each class. We got 100% accuracy for bar charts, pie charts and line graphs. The model is correctly identifying algorithm images with 93% accuracy, equations with 100% and network images with 97.5% given by fig.6. Our model is compared with the accuracy of SVM as claimed by author [9] and gives better accuracy for all the image classes. The performance of our model is also evaluated using a confusion matrix. It gives the count of images in one class, misclassified as another class and the diagonal gives the precise categorisation. Fig.5 shows the performance of the classifier. We get the following results from the confusion matrix.

- All the images in bar charts, pie charts, equations and algorithm gives 100% correctly classified.
- One line chart, is misclassified as scatter plot and two scatter plots are misclassified as a line chart. The scatter plots which got confused is very similar to the dotted line graph.
- One of the network images is misclassified as an equation, this image from the internet vary from others and it has more textual data in it.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 32, 198, 198)	320
max_pooling2d_1 (MaxPooling2	(None, 6, 39, 198)	0
conv2d_2 (Conv2D)	(None, 32, 37, 196)	1760
max_pooling2d_2 (MaxPooling2	(None, 6, 7, 196)	0
conv2d_3 (Conv2D)	(None, 32, 5, 194)	1760
max_pooling2d_3 (MaxPooling2	(None, 6, 1, 194)	0
dropout_1 (Dropout)	(None, 6, 1, 194)	0
flatten_1 (Flatten)	(None, 1164)	0
dense_1 (Dense)	(None, 128)	149120
dense_2 (Dense)	(None, 8)	1032

Total params: 153,992
Trainable params: 153,992
Non-trainable params: 0



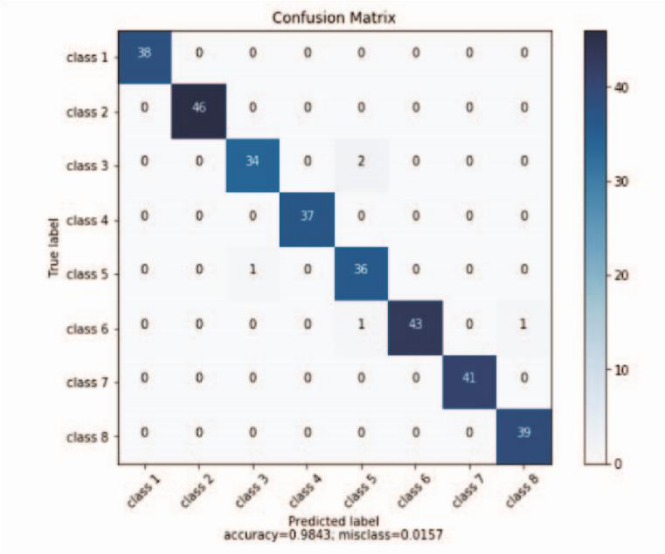
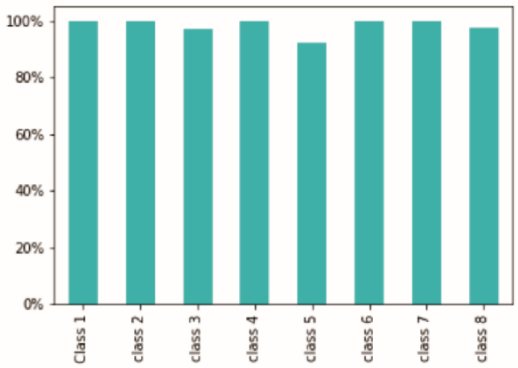


Fig. 5. Performance of the classifier as confusion matrix. class 1: Vertical bar graphs, class 1 : Horizontal bar graphs, class 3: Line charts , class 4: Pie charts, class 5: Scatter plot, class 6: Equations, class 7: Algorithms, class 8: Networks



IV. RESULTS

V. SUMMARY AND FUTURE WORK

Our model gave better accuracy in identifying the types of Vertical bar graphs, class 1 : Horizontal bar graphs, class 3: Line charts , class images in a document. It was tested on an image from news- 4: Pie charts, class 5: Scatter plot, class 6: Equations, class 7: Algorithms, class 8: Networks paper which consists bar graph and line charts and resulted

TABLE III
COMPARISON OF PROPOSED METHOD WITH METHOD [9]

Types	Method [9]	Proposed Method
Bar charts	90%	100%
Line charts	76%	97%
Pie charts	83%	100%
Scatter plots	86%	92%
Surface plots	84%	-
Equation images	-	100%
Algorithm images	-	100%
Network images	-	97.5%

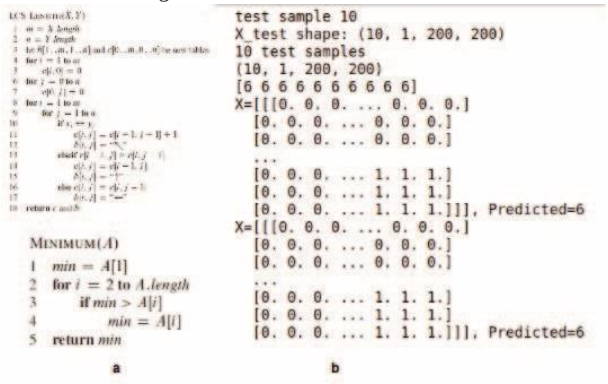


Fig. 7. a: An instance of input samples belonging to class 6, b: All the test samples correctly predicted as class 6

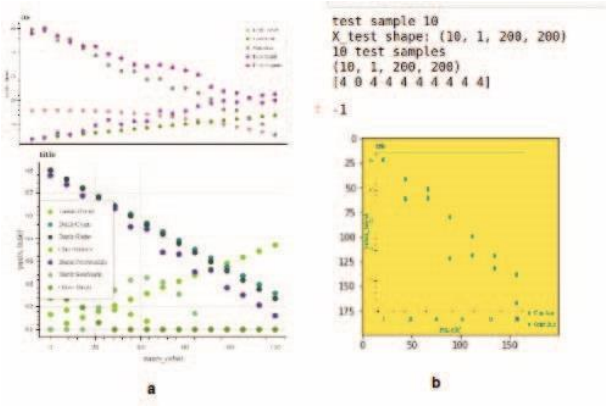


Fig. 8. a: An instance of input samples for scatter plot ie, class 4, b: One mislcassification, the second image misclassified as line plot is shown

in predicting correctly. The reason for misclassification in the testing is due to the difference in input images from train images as some of them were using large or different fonts. There were variations in the textual content in the pictures from the internet. We want to improve the accuracy of the classes which showed misclassification and also localize the figures in a document. As a next stage, will extract the textual information in these images and also analyse the underlying data of these charts and thus produce a detailed description of the image as alternative text/audio to help the blind. Further, extend this to educate the blind through voice output.

REFERENCES

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. Introduction to Algorithms, Third Edition (3rd ed.). The MIT Press.

[2] Yu, Wai, and Stephen Brewster. ”Evaluation of multimodal graphs for blind people.” Universal Access in the Information Society 2, no. 2 (2003): 105-124.

[3] Moustakas, Konstantinos, Georgios Nikolakis, Konstantinos Kostopoulos, Dimitrios Tzovaras, and Michael G. Strintzis. ”Haptic rendering of visual data for the visually impaired.” IEEE MultiMedia 14, no. 1 (2007): 62-72.

[4] Alison Bert, DMA and Lisa Marie Hayes. ”Making charts accessible for people with visual impairments” 2017 . [Online]. Available: https://www.elsevier.com/connect/making-charts-accessiblefor-people-with-visual-impairments. [Accessed: 14- Mar-2019]

[5] Elzer, Stephanie, Edward Schwartz, Sandra Carberry, Daniel Chester, Seniz Demir, and Peng Wu. ”A Browser Extension for Providing Visually Impaired Users Access to the Content of Bar Charts on the Web.” In WEBIST (2), pp. 59-66. 2007.

[6] Lazar, Jonathan, Aaron Allen, Jason Kleinman, and Chris Malarkey. ”What frustrates screen reader users on the web: A study of 100 blind users.” International Journal of human-computer interaction 22, no. 3 (2007): 247-269.

[7] Siegel, Noah, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. ”FigureSeer: Parsing result-figures in research papers.” In European Conference on Computer Vision, pp. 664-680. Springer, Cham, 2016.

[8] Kahou, Samira Ebrahimi, Vincent Michalski, Adam Atkinson, kos Kdr, Adam Trischler, and Yoshua Bengio. ”Figureqa: An annotated figure dataset for visual reasoning.” arXiv preprint arXiv:1710.07300 (2017).

[9] Balaji, Abhijit, Thuvaarakkesh Ramanathan, and Venkateshwarlu Sonathi. ”Chart-Text: A Fully Automated Chart Image Descriptor.” arXiv preprint arXiv:1812.10636 (2018).

[10] Kita, Yui, and Jun Rekimoto. ”Prediction of importance of figures in scholarly papers.” In 2017 Twelfth International Conference on Digital Information Management (ICDIM), pp. 46-53. IEEE, 2017.

[11] Mirri, Silvia, Silvio Peroni, Paola Salomoni, Fabio Vitali, and Vincenzo Rubano. ”Towards accessible graphs in HTML-based scientific articles.” In 2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC), pp. 1067-1072. IEEE, 2017.

[12] Elzer, Stephanie, Edward Schwartz, Sandra Carberry, Daniel Chester, Seniz Demir, and Peng Wu. ”A Browser Extension for Providing Visually Impaired Users Access to the Content of Bar Charts on the Web.” In WEBIST (2), pp. 59-66. 2007.

[13] Talib, Muhammad Nabeel, Cai Shuqin, Muhammad Abrar, and Muhammad Sheraz Shafiq. ”E-business access for blinds: A semantic approach.” In 2009 International Conference on E-Business and Information System Security, pp. 1-4. IEEE, 2009.

[14] Clark, Christopher Andreas, and Santosh Divvala. ”Looking beyond text: Extracting figures, tables and captions from computer science

papers." In Workshops at the Twenty-Ninth AAAI Conference on Artificial
Intelligence. 2015.

[15] Prasad, V. Shiv Naga, Behjat Siddiquie, Jennifer Golbeck, and Larry S. Davis. "Classifying computer generated charts." In 2007 International Workshop on Content-Based Multimedia Indexing, pp. 85-92. IEEE, 2007.

[16] Savva, Manolis, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. "Revision: Automated classification, analysis and redesign of chart images." In Proceedings of the 24th annual ACM symposium on User interface software and technology, pp. 393402. ACM, 2011.

[17] Clark, Christopher, and Santosh Divvala. "Pdffigures 2.0: Mining figures from research papers." In 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), pp. 143-152. IEEE, 2016.

[18] Al-Zaidy, Rabah A., and C. Lee Giles. "Automatic extraction of data from bar charts." In Proceedings of the 8th International Conference on Knowledge Capture, p. 30. ACM, 2015.

[19] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436.

[20] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15, no. 1 (2014): 1929-1958.