

Classification Of Visualization In Scientific Literature

Masterarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

Lehrstuhl für Intelligent Systems und Lehrstuhl für Data Science
der Fakultät für Informatik und Mathematik
der Universität Passau

Name:	Arnold Azeem
Matrikelnummer:	79176
Fachbereich:	Informatik
Studiengang:	Master Informatik
Erstprüfer:	Prof. Dr. Christin Siefert
Zweitprüfer:	Prof. Dr. Michael Granitzer
Date:	June 19, 2018

0.1 Abstract

Distinct visualization techniques are used in scientific research publications to summarize large amount of data and also represent a variety of data. These visualizations help to communicate complex information and support the arguments being presented in the publication in a way that is easy to understand and follow. These figures tend to reveal trends, patterns or relations that might have otherwise been difficult to grasp using only text. It is therefore relevant that we extract the data from these visualizations since the extracted data can be used for validating the publication or presenting the data in another form for a different audience. In this context, classifying these visualizations is the initial step since, there is a variety of visualizations and each one is processed in a specific way. It is only after classification that extracting of raw data from these visualizations can be acquired for other tasks. This thesis presents an approach whereby real world data is used to create four types of plots (scatter plots, bar charts, line charts, and box-plots) and random plots also of the same kind from the Internet are added together and used to train and evaluate a CovNet model to be able to classify these plots.

Contents

0.1	Abstract	2
	List of Figures	3
	List of Tables	4
1	Acknowledgements	5
2	Introduction	6
2.1	Overview	6
2.2	Motivation	8
2.3	Objective	9
2.4	Research Significance and Contribution	10
3	Background and Related Work	11
3.1	Model Based Approaches	11
3.2	Machine Learning Based Approaches	12
	The Brain	13
3.2.1	Input data format	13
3.2.2	Convolutional Neural Network	14
	Convolutional Layer	15
	Activation Functions	15
	Pooling Layer	16
	Dropout Layers	16
	Fully Connected Layer	17
	Training process	18
	Machine Learning Classification Algorithms to Recognize Chart Types in Portable Document Format(PDF) Files	18
	Architecture proposal for data extraction of chart images using Convolutional Neural Network	20
	Chart classification by combining deep convolutional networks and deep belief networks	21

4	Approach	23
4.1	Dataset	23
4.1.1	Dataset for Matlab	24
4.1.2	Dataset for R	24
4.1.3	Dataset for Python	25
4.1.4	Dataset for Java	25
4.2	Creating Plots	26
4.2.1	Random Charts	26
4.2.2	Training, Validation and Test set	30
4.3	The Architecture	31
	Convolutional Layer	32
4.3.1	Dropout	32
4.3.2	Fully connected layers	32
5	Experiments	34
6	Summary	35
7	Bibliography	36

List of Figures

2.1	Seletecting the appropariate chart for visualizing data	7
2.2	A box plot with a legend, x-axis, and y-axis from [1]	8
2.3	Overview of process to obtain raw data from scientific papers. The red outline indicates the work in this thesis.	10
3.1	How a computer perceives an image- source [2]	14
3.2	ReLU unit compared to the tanh unit from Krizhevsky et al. paper [3]	16
3.3	Left: shows a standard neural net. Right: shows a neural net- work where dropout has been applied on. Crossed units are dropped units. figure from Krizhevsky et al. paper [4]	17
4.1	Scatter plots	26
4.2	Example Bar Charts	29
4.3	Example Line Charts	29
4.4	Example Box Plots	29
4.5	Example Chart images from Po-shen et al. and Junior et al. . .	30
4.6	Steps involved in building the Classification model	31
4.7	Architecure of CovNet used to train our model	32

List of Tables

3.1	Details on Dataset from [5]	19
3.2	Number of Train and Test Dataset collected	20
3.3	Comparing Results of Proposed Framework from (Liu et al. [6])	22
4.1	Names of datasets used in each plotting program	25
4.2	Overview of the varied parameters and libraries used for creating plots in the different plotting programs	27
4.3	Overview of the number of plot and different varieties plots in the different plotting programs	28
4.4	The following table shows different layers, parameters and computation units used.	33

1 Acknowledgements

2 Introduction

"A picture is worth a thousand words" even though a widely used phrase stands to be very true especially when dealing with scientific literature. This is due to the fact that information and data is easily understood when visualized, for this reason most scientific publications tend to present their data using tables, plots and pictures. Identification, extraction and interpretation of this visualized data for purposes of; validation of results by another researcher, or presentation of the data to another audience is an area in machine learning which is still found wanting.

The aim of this thesis is to develop a reliable system for the purpose of recognizing chart images by using example chart images created with real world data. Image processing and machine learning methods are the main constituents of this thesis. The initial section in this chapter gives an overview of chart image classification and the complications and difficulties faced in chart image classification. The other parts of this paper are organized in the follows: In the other sections of this chapter the research motivation, objectives, contributions and an outline of the proposed framework in this thesis is presented. Finally, the structure of the rest of this thesis is illustrated in Section 1.6

2.1 Overview

A basic description of a chart is a graphical representation of data, where the data is represented by some type of symbols. These symbols could be rectangles in the case of histograms or bar charts, slices that make a circle in the case of a pie chart, and dots, stars, diamonds and an asterisk in the case of a scatter plot. Charts are often used to visualize complex data for easy understanding and interpretation, to find patterns and trends in data. Charts can usually be read more quickly than textual data, they are used in almost every field especially in science, marketing, and maths. Charts are mostly created with the help of a computer but can also be created by hand. There are a variety of charts and therefore, to present your data in the right way, the right type of chart must be employed. For example bar charts are best used when

comparing data, line charts works better when we want to find trends in data and scatter plots are used when we want to find relationships in data. Figure 2.1 which was inspired by the figure from [7] shows which charts we should choose depending on what we would like to describe in some acquired data. For example if we want to show a relationship between two variables in some data, we should a scatter plot as seen in the figure.

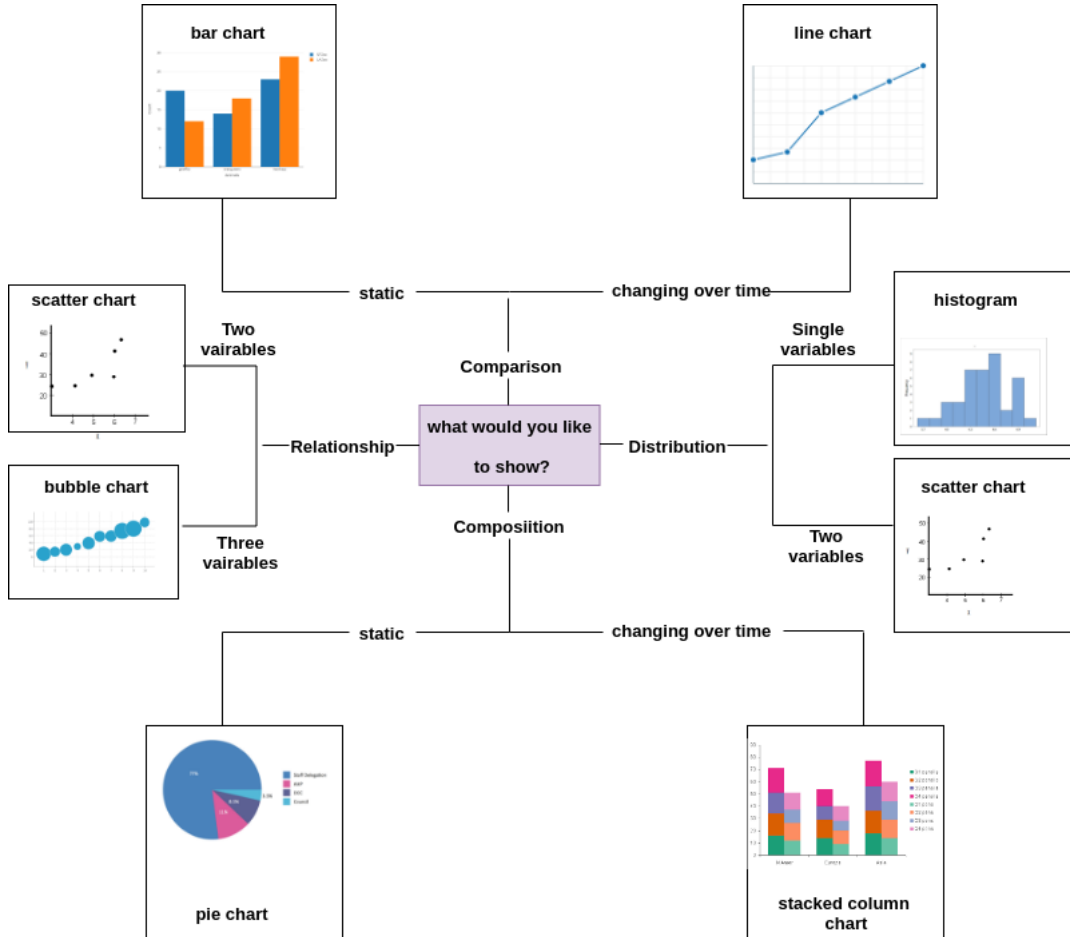


Figure 2.1: Seletecting the appropariate chart for visualizing data

Charts usually appear in different forms, nonetheless there are some features that are mostly present in every chart. Charts are mostly graphical with a little bit of text used to annotate the data. One important use of text visible in most charts is to give the chart a title. The title describes what the graphic represents and is usually situated above the main graphical part. Text is used less because the brain processes visual information sixty thousand faster [8]. Most graphs have one line at the side and another usually at the bottom. These lines if used are called axes of the graph. The one at the bottom is called the x-axis, and the other line the y-axis. Each graph usually indicates numerical or periodic sequence shown by periodic graduations. There are also texts usually outside or besides the axis describing the measurements. Some

charts include what we call a legend, the legend is usually present in situations where there are multiple variables to be described. Figure 2.2 shows a typical chart that has the features mentioned above. The legend showing the two variables female and male represented in the chart, the label on the left side annotating the cholesterol levels with the different measurements of the cholesterol shown on the y-axis, and to the bottom categorical measurements shown on the x-axis. Finally the title, that is the text describing the whole chart shown at the top.

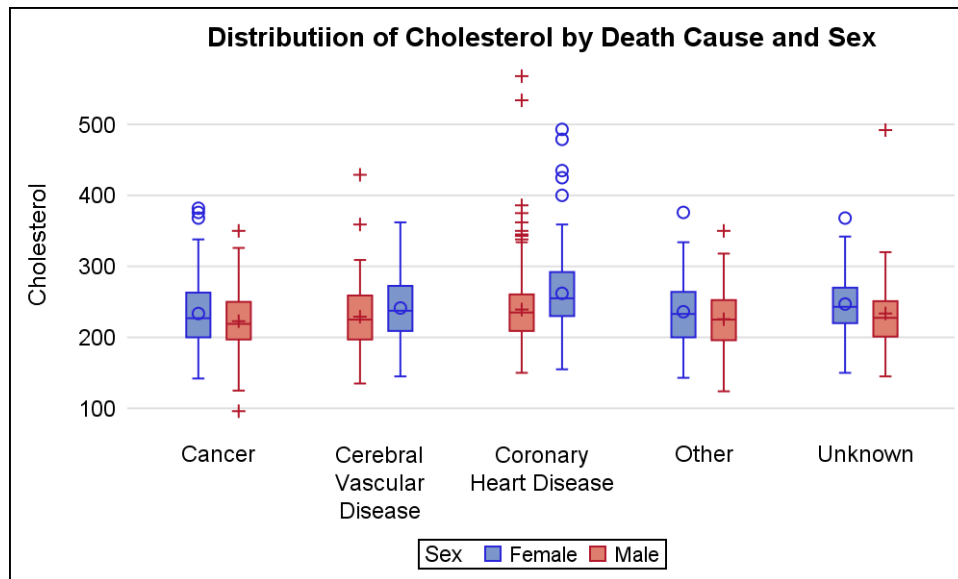


Figure 2.2: A box plot with a legend, x-axis, and y-axis from [1]

Despite most features being present in most charts, there are some charts that are totally different from the rest. For example a pie chart does not have an x-axis or y-axis, also the designers of a particular chart have a choice of visual encoding and styling, this results in a wide variety of charts, even in charts of the same type. To be able to identify the different types of charts will mean having all the wide range of the different visual encodings and styles of these charts.

2.2 Motivation

Complex data is better explained in scientific documents with the aid of visualizations. These charts present complex data in an easy to understand format as compared to the textual content. Chart images are a very important part of scientific documents. Data is ever growing and sometimes complex, and therefore using figures and diagrams to interpret and represent this data cannot be undermined. Chart images provide a way to easily give insight to research findings, which would have otherwise been more complex relying on only textual data. The data which these visualizations contain when extracted play

an important role in events where another researcher wants to verify the work of the publisher. The extracted data can also be used to develop other visualizations in situations where the paper needs to be presented to a different audience with a distinct background as opposed to the audience for which the visualizations were created initially. Another reason for extracting this data is that, when comparing two plots, looking at the raw data, gives us more insight and consequently helps in better decision making as compared to just looking at the figures. Since each plot will be processed differently to extract the raw data, it is very relevant that we can distinguish one plot from another. For this reason, there has been a growing interest in the chart analysis area and quite a number of techniques have been developed. In spite of this growing interest, there has been little groundbreaking results achieved due to different variations in appearance of plots [6]. For example, Manollis Savva, et al [9] proposed a model to classify charts using extracted low-level features and textual features. After extracting the features, a Support Vector Machines (SVMs) classifier is used for the classification step. This method was limited since most charts contain the same type of features like axes, grid lines, and legends. In V. Shiv Naga Prasad's work [10] classification was based on using features based on the shape and spatial relationships of their primitives. This work was limited due to the inconstancy in which data in most charts can be depicted. However in the paper of Alex Krizhevsky et al. [3] a deep CovNet which combines both feature extraction and classification was used and this technique achieved very good results.

2.3 Objective

The aim of this thesis is to develop algorithms using image processing and machine learning methods for the purpose of chart image classification. Different approaches will be considered in order to choose the most favorable one.

The goal of this thesis is to try to identify or recognize chart images with the highest accuracy. To achieve this goal we have to answer the question:

How Well Can We Classify the Four Different Types of Plots (Line-Charts, Bar-Charts, Scatter-plots, and Box-plots) in Scientific Literature?

To answer this question, the following objectives are set.

- obtain raw data gathered from real world occurrences.
- create a dataset of chart images with the collected data. Using different plotting programs (Python, Matlab, R, and Java) and different libraries supported by these plotting programs.

- train and evaluate a machine learning model that can recognize unseen charts with high accuracy. The four classes of plots are box plots, line charts, scatter plots and, bar-charts.

2.4 Research Significance and Contribution

Figure 2.3 shows the significance of this work and what it will lead to. This thesis focuses mainly on four chart images. These plots are scatter-plots, bar-charts, line-charts because they are 3 of the most popular image charts and box-plots because it shows skewness and unusual observations in a dataset and this information is especially relevant when dealing with large data, which is mostly the case in machine learning. The first part of the diagram involves, obtaining the four different types of plots mentioned earlier, after which we then label our plots and train a neural network model to be able to classify with high accuracy any of the four plots if shown to our model, then finally the raw data can be extracted from the detected plot using other processes and techniques. The focus of this thesis however, is shown in the figure with red dotted lines, which entails, gathering a dataset of chart images, labeling them, training the model and classifying the plots.

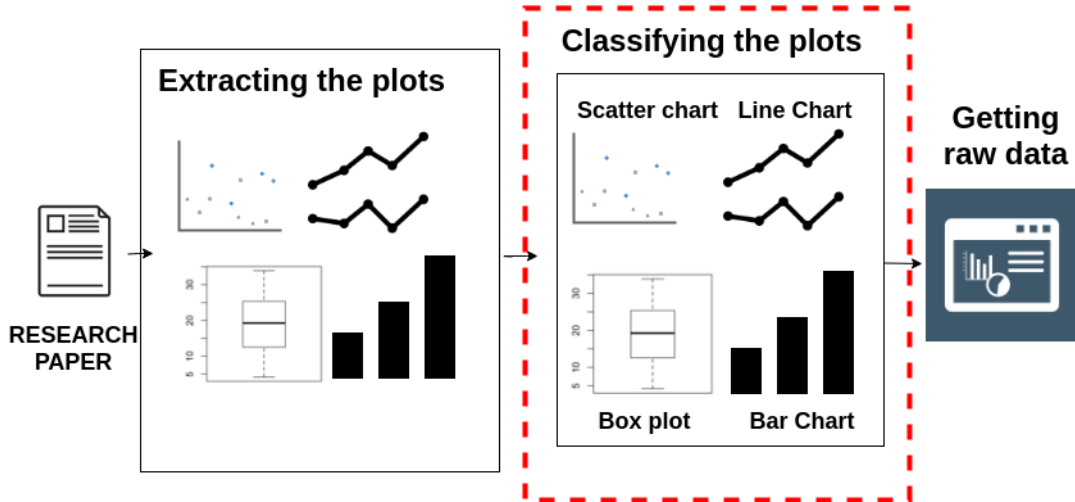


Figure 2.3: Overview of process to obtain raw data from scientific papers. The red outline indicates the work in this thesis.

3 Background and Related Work

Classification involving charts images has been a topic of huge interest in recent times. This is due to the fact that, most scientific publications if not all, are embedded with these charts as a way of conveying complex research findings in more visualizable and easy to understand format. As a result of this popularity, chart images have gained, different techniques have over time been used for classification of charts and these techniques have evolved over the years. The techniques have been inspired by techniques from image processing, raster to vector conversion and layout analysis. These techniques can be put into two categories; Model-Based Approaches and Machine Learning Based Approaches [11]. In the next section, the methods and techniques used in this thesis, and techniques that inspired us to choose these techniques are explained in details.

3.1 Model Based Approaches

The following is written with inspiration from Boyle et al. work [12]. Numerous chart images even of the same type, are very different in sense of their context structure. This is as a result of the variability of positions of the context structure of a chart. There is exists no real standard for all charts to follow, content like; legends, axes, grids etc., have no fixed positions and sometimes are not even present in some charts. As a result of this non standardization of charts, classifying becomes a very difficult task. As a results this approach tries to uses the structure of the charts. The model-based approach is divided into two steps. Firstly, a number of predefined object classes are created with abstract models and then an image is matched with the created model. For example, a model for a 2D/3D pie chart consists of line segments (radii and circular/elliptical arcs), and these line segments are used to create the model, however, just using these line segments wont be enough, so some constraints are introduced into the model, example of constraints of a pie chart include; the center of the pie chart is where all radii meet; all 2D pie charts have radii of equal length; arcs mainly form parts of the same 2D pie chart circle, or if its a 3D then the arcs form part of the ellipse. After this a goodness-of-fit is introduced to help measure the discrepancy between the image and the condi-

tions for the model. For example, a goodness-of-fit of a pie chart is as follows: difference in length of radii; difference in how the arcs are curved; distance between a center of a circle and that of a radii. When a chart is presented for classification, the edges are detected, thinned, linked, vectorized, extracted and compared to the various edge models created and the best match is selected. After this, the good-to-fit criteria is used to measure the difference between the edges of the model and the image. Finally, a voting is performed base on the value of the good-to-fit results. The model-based approach has some drawbacks though. The main drawback is that human intervention is needed to develop the various models. This could be a cumbersome and time consuming procedure. Due to the limitation of this approach a machine learning approach was introduced to handle some of the limited of this approach.

3.2 Machine Learning Based Approaches

This method was developed in recent times, and most works in area of text mining and image processing employ this approach. This method involves using handcrafted features extracted from the charts for classification task. In Zhou and Tan’s [13] work, features like legend, x-y-axis title, chart title, and values of the bar of a bar-chart were extracted for purposes of classifying a bar-chart. In Shao and Futrelle’s [14] work, graphical elements of the charts like colors, tick marks on an axis and data point markers are the handcrafted feature used for classification. Another instance where handcrafted features were used was in Inokuchi et al. work [15]. In this work regularly appearing substructures of chart are extracted and used as features for classification. After the feature acquisition step, the obtained features are then represented in vector form for the classification step. The features are stored in a matrix vector-like structure. Algorithm 1 shows an example of the structure in which the features are stored. This structure has multiple columns containing the features and one column indicating the target or label of the chart. This vector-like structure is then fed as input into a machine learning technique for the model training step. In Karthikeyani and Nagarajan’s paper [5] after the features were obtained SVM Classification ¹, MLP Classification ², and KNN Classification³ were used to create a function that maps the set of features to a predefined label. This technique achieves very good results but, the drawback is the over reliance on handcrafted features. This drawback makes it difficult when large amount of data consisting of a variable context structure is involved, and this situation is evident in most cases. Recently, however CovNets

¹<http://www.statsoft.com/Textbook/Support-Vector-Machines>

²http://www.iro.umontreal.ca/~bengioy/ift6266/H12/html/mlp_en.html

³<https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

(this abbreviation stands for Convolutional Neural Networks and will be used a lot through this work) ⁴, a machine learning technique which removes the drawback of using handcrafted features by learning important features anywhere in the image by itself. CovNets performance on image classification keeps on improving and in recent times, results are being compared to human performance levels. CovNets were used in this work, in the next section they will be explained into details, and is inspired by an article written by Daphne Cornelisse [16].

The Brain

The following is . As human beings we identify objects around as all the time with minimal effort. We identify objects based on what we previously leaned. We see objects, label, recognize patterns, and make predictions subconsciously everyday. How are we able to do this effortlessly? Even though the whole process starts with the eye, most of the process happens in the part of the brain called the primary visual cortex. When we see an object, light receptors send signals via the optic nerves to the visual cortex and sense is made of the perceived object by this organ. The deep neural connections in the brain plays a major role in remembering and identifying objects.

3.2.1 Input data format

This section was inspired by Nikhil B's article [17]. Just like the way the human eye needs to perform some preprocessing steps before an image signal reaches the brain, an image recognition system can better identify images if some preprocessing steps were performed on the input images. Below some effective preprocessing steps that go a long way to improve image recognition are explained:

- **Uniform aspect ratio:** Neural networks usually take as input square shaped images, all images should consequently have the same size and aspect ratio. Images should for this reason be checked and cropped to meet this requirement appropriately. One important thing we should consider when cropping is to keep the middle part as much as possible.
- **Normalizing image inputs:** This step is mainly performed to ensure that there is uniform distribution in the image pixels data distribution. This in the long run accelerates convergence during the process of training.

⁴<https://deeplearning4j.org/convolutionalnetwork>

- **Dimensionality reduction:** Depending on the type of experiment being performed, the RGB channels of a colored image can be collapsed and converted to a gray-scale channel. This can result in a faster training time.
- **Data augmentation:** This step is also very dependent on the type of experiment being done. Variant of the input image are acquired by augmenting the existing dataset. This includes scaling, rotation etc.

3.2.2 Convolutional Neural Network

Just like how human beings learn to recognize and label objects, we need to show an algorithm a huge number of images for it to be able to identify an image it has never seen before. Computers unlike humans perceive images as a 2D array of numbers, known as pixels. A CovNet has neurons just like the visual cortex and these are organized into layers. Each layer tries to identify a part of the input image. For example one layer identifies the edges, another identifies the curves and so on. This is done by employing the spatial relation that exist in the pixel of an image. All the layers are connected to each other. There are four main layers in the CovNet:

- Convolution
- Non Linearity (ReLU)
- Pooling or Sub Sampling
- Classification (Fully Connected Layer)

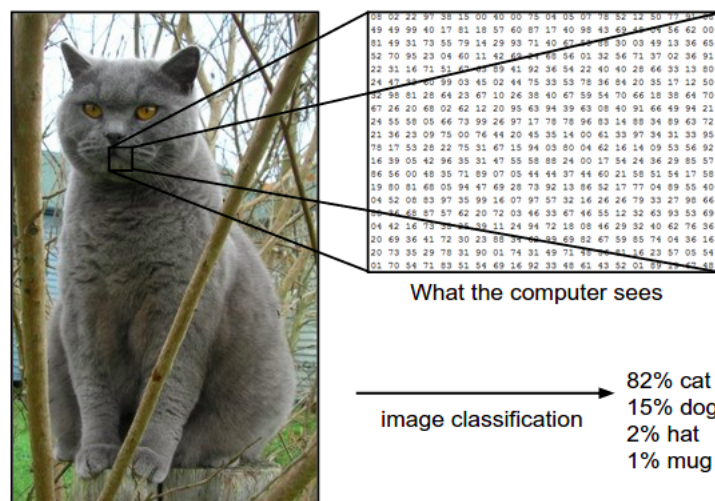


Figure 3.1: How a computer perceives an image- source [2]

Convolutional Layer

The Convolutional layer will be explained using the following example, let consider the image as a magical cake of height and width 48x48, and there is a cake cutter of size say 5x5. The cutter is used to cut the cake from the top left. The cutter is referred to as a filter in machine learning. The cutter in this case is also an array of numbers called weights and for the maths to work the cutter must have the same depth as our cake (5x5x3). Lets first consider the cutter being in the first position ie. the top left of the cake. The values in the cutter are multiplied with the values of the cake (element wise multiplications). These multiplications are all summed up and result in a single number. This single number is only for the first part and this process is repeated through out the whole cake (Next step would be moving the cutter to the right by 1 unit, then to the right again by 1, and so on) keep in mind its a magical cake so you can cut a part more than once. After using the cutter on the whole of the magical cake we result in a new cake of size 28x28x1, the results is called a feature map. The filters are low level feature identifiers (straight edges, simple colors, and curves).

Activation Functions

The activation functions are a very important part of a CovNet, they are mostly added to the output end of a Convolutional layer. It usually maps the output values between 0 and 1 or -1 and 1. The activation functions can be divided into two categories; linear and non-linear functions. The non-linear functions are the most used because most of real world data are non-linear. There 3 main types of this function that are mostly used. They are the sigmoid, tanh and ReLu. The sigmoid map the output value between 0 and 1, the tanh maps the output between -1 and 1 and finally the ReLu the most used activation function. The Rectified Linear Units (ReLU) is found in all layers during the Convolution phase, this layer implements element-wise nonlinearity to our Convolutional layer, this just means it helps to handle situations where the relation between the input values and the CovNet output is non-linear. The ReLU has a function $f(x) = \max(0, x)$ which means if you give it a value x, it will return 0 if x is negative and will return the value itself if its positive. We used ReLU mainly because it speeds up the training process significantly and it does not saturate(the gradient is very small, if the input is elevated or small) unlike the tanh and sigmoid [18]. Figure 3.2 shows a plot from Krizhevsky et al. [3] paper shows that there is a 6x advancement in convergence with the ReLU unit compared to the tanh unit. The bold line represents the ReLU and tanh is the dashed line.

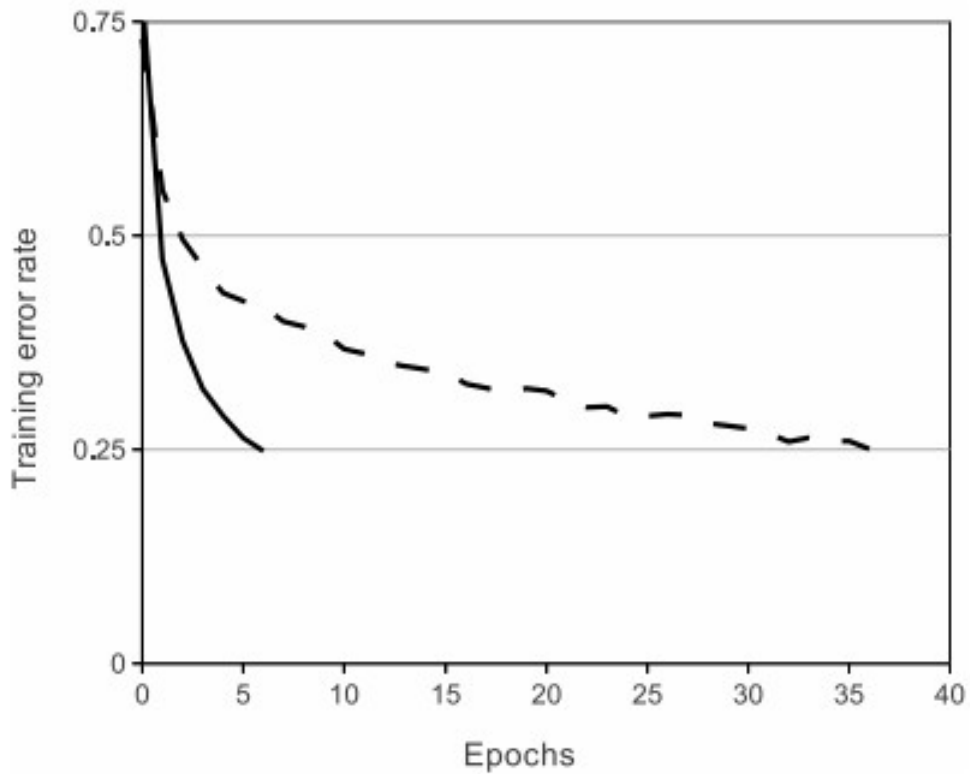


Figure 3.2: ReLU unit compared to the tanh unit from Krizhevsky et al. paper [3]

Pooling Layer

The pooling layers basically have two main functions, first one is that it helps the CovNet to locate features regardless of which part of the image it is located. This results in the model being robust against small changes in position of the features of the images. The second function is that it also helps to reduce the size of the feature map. Therefore computations in the futures layers are relatively less complex. One way of performing pooling is using max pooling technique other less used techniques include Average and sum pooling, the idea in max-pooling is like sliding a window through the feature map, the windows fills a number of arrays in the feature map therefore, you pick the largest among all the numbers and disregard the rest of the numbers. We employ this technique in the first, second and last Convolutional layers of our network.

Dropout Layers

The following is inspired by an article on Medium [19]. This layer is used for regularizing your network. The dropout is used in our CovNet to prevent over-fitting. The Fully connected layers eventually handles more parameters and therefore neurons become co-dependent on one another, and this leads to over-fitting during the training phase. When dropout is implemented in a network, it does not use all neurons during a particular forward or backward

pass during training but, chooses random neurons at a specified probability for each pass. Figure 3.3 shows how a dropout is different from a normal neural network.

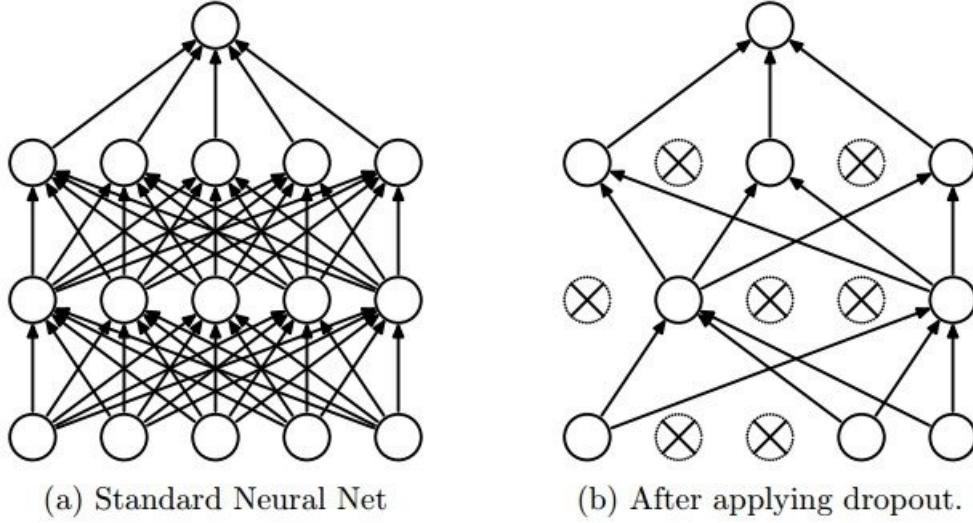


Figure 3.3: Left: shows a standard neural net. Right: shows a neural network where dropout has been applied on. Crossed units are dropped units. figure from Krizhevsky et al. paper [4]

Fully Connected Layer

Fully connected simply implies that every neuron from the previous layer is connected to every other neuron in the next layer. The output of this layer are high level features of the image presented to the network. The output of this layer is a probability distribution of all the classes in the network and this is ensured by using an activation function, usually a softmax activation function. We used a softmax activation as opposed to other activation functions because, it handles low invigoration in images (say blurry charts) of your neural net with rather uniform distribution and to high invigoration in images (say large numbers, like sharp images) with prediction probabilities between 1 and 0. To put in simple terms, the Convolutional and pooling layers act as feature extractors and the last layer, the fully connected layer act as the classifier.

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \quad (3.1)$$

The equation 3.1 is a softmax function, (z) represents a vector, so the function takes (in z) and crushes it to a probability value between zero and one [2].

Training process

This section is inspired by a blog written by ujjwalkarn [20]. The process by which a CovNet performs classification is summarized below.

- All parameters (weights, filter, and biases) are initialized randomly.
- The network is then given as input the image. The forward pass is then performed (convolution, relu, pooling and fully connected layer), the output is a probability of each class.
- here the error in the output is calculated.

$$TotalError = \sum \frac{1}{2}(targetprobability - outputprobability)^2$$

- This step reduces the total error by performing back-propagation, the gradients of the error with respect to the weights are calculated. This is used to update parameters and therefore results in reducing the total error.
- In the final step, steps 2 to 4 are repeated for all images in the training dataset. The best parameters are then stored for the prediction of a new image given to the network.

In the rest of this chapter, three related works that performed chart image classification, with very unique methods and also achieved good results are described in details.

Machine Learning Classification Algorithms to Recognize Chart Types in Portable Document Format(PDF) Files

Karthikeyani and Nagarajan's work [5] focuses on classifying charts found in pdf files. The steps employed for the classification task involves extracting texture features from the charts and then, using a machine learning technique for the classifying phase. The dataset consists of images extracted from various pdf's. The dataset is made up of 155, 256*256 RGB images in total. Table 3.1 shows the details of the dataset used. The technique used involved extracting handcrafted features from the chart image and then, using these features as input for a model to be trained and tested. For the feature extraction phase, Gray Level Co-Occurrence Matrix (GLCM) is employed. GLCM is a technique which uses co-occurrence matrix to extract texture features of an image with the use of statistical equations. GLCM was used to extract eleven

features, these included: area, median, minimum and maximum intensity, contrast, homogeneity, energy, entropy, mean, variance, standard deviation and correlation. These extracted features are correlated to the pixels of the image. These extracted features are then stored using a 2-dimensional matrix vector data structure. This structure has thirteen (13) columns and 'x' rows, where x is the size of the dataset. The features are stored in twelve (12) columns and the label is stored in the thirteenth column. The features and labels were stored in the below structure 1.

Chart Type	No of Charts	Chart Type	No of Charts
2D Bar chart	40	Doughnut 2D	7
3D Bar chart	16	Doughnut 3D	11
2D Pie chart	13	Line	35
3D Pie Chart	20	Mixed Chart	13

Table 3.1: Details on Dataset from [5]

Algorithm 1 This is the structure used to store the features

```

Struct FeatureVector {
float feature1; float feature2;
float feature3; float feature4;
float feature5; float feature6;
float feature7; float feature8;
float feature9; float feature10;
float feature11; float feature12;
int target;
}

```

This vector like structure is then fed as input into three classifiers SVM, MLP neural network, and K-Nearest Neighbor. These classifiers are then trained and a classification model is formed for the recognition step. After this, to see if the model works well a test set consisting of new records is fed into the model for the model to predict their labels. Three metrics were used to check the performance of the model, these metrics are; error rate, classification accuracy and speed of classification. The error rates obtained were as follows; MLP 0.30, K-NN 0.22 and SVM 0.23. For the accuracy metrics were as follows; KNN (78.06%), MLP (69.68%) and SVM (76.77%) and finally the speed metric. The speed metric which sum of training and test time results are; MPL was the slowest with 8.38, followed by SVM with 0.31secs and KNN with 0.26secs. Even though these results were very good, the paper proposed extracting features which are related to shape and curves since these features will carry more unique information to distinguish charts.

Architecture proposal for data extraction of chart images using Convolutional Neural Network

Our work is inspired by De Freitas et al. [21], proposed a way to extract the wealth of information contained in different visualization techniques. The paper talks about two main stages of accomplishing this task. Firstly, classification of the charts is done since it allows a different variety of chart to be detected automatically allowing the next step, which is the extraction of data from the classified plots. The paper, however, focuses on the first step, classification of charts. In this paper, a Convolutional Neural Network is used for the classification task. The Convolutional neural network encapsulates the characterization and classification processes during its learning process, unlike other techniques. The dataset used for this task were searched for and downloaded from Google image search. Table 3.2 shows the chart types which were collected and the number of train and test sets which the respective charts were divided into.

Chart Type	Test	Train
Area Chart	50	555
Bar Chart	50	657
Line Chart	50	489
Map	50	476
Pareto Chart	50	261
Pie Chart	50	361
Radar Chart	50	454
Scatter Chart	50	552
Table	44	236
Venn Diagram	48	304
Total	498	4345

Table 3.2: Number of Train and Test Dataset collected

For the classification, a variant of CovNet called LeNet-based CovNet model is used. The model was implemented using Tensorflow⁵, LeNet-based CovNet has an architecture which is comprised of 3 convolutional layers, followed by a fully connected layer. The model is trained in a way that the dataset is divided into mini-batches, samples of fixed sizes(100) are selected and fed into the CovNet, as a result of this process the model becomes robust since it learns to generalize from the different min-batches which are fed into the

⁵<https://www.tensorflow.org/>

model. Also, all the images are converted to JPG and resized to 224x224x3, that is, 224 pixels of height, 224 of width and 3 layers of output. The other parameters used were 1000 epochs and a learning rate of 0.003. The accuracy at the end of the training process was 70%.

Chart classification by combining deep convolutional networks and deep belief networks

In another paper by Liu et al. [6], a new approach was proposed for the process of chart classification. The process involves using CovNet to extract deep hidden features of charts and then deep belief networks then use the extracted features to predict the labels of the charts. Due to a difficulty in acquiring a large number of charts as training data, natural images were first used to train the model and later the model was fine-tuned with just over 5,000 collected charts. The types of charts collected were pie charts, scatter charts, line charts, bar charts, and flowcharts. The architecture of the CovNet is made up of five Convolutional layers and two fully-connected layers and then an output layer. The preprocessing steps for the images involve down-sampling them to 256 x 256 x 3, after which each is cropped to a size 227 x 227 from the center and its horizontal flip are extracted as the input of the CovNet, other parameters used for the CovNet include a learning rate that starts with 0.01 initially and is then decreased by a factor of 0.1 after every 100k iterations, the weight decay parameter was set at 0.0005 and a dropout rate of 0.5. This results in an output of a 5-way softmax which produces the distribution over the 5 class labels and this is used as input for the deep belief network. The deep belief network architecture has three hidden layers, whose dimensions are 5000, 500 and 2000. This results in a softmax predicting the probability distribution over the 5 categories of charts as output. The training process was done with 4000 randomly selected images and the rest were used as test set. The accuracy of the model after the evaluation was 75.4%. Table 3.3 shows the results after the training was done without deep belief networks but pre-trained with the natural images and finally the training done with only the chart dataset but with deep belief networks.

Chart	CovNets	CovNets+DBN without pre- training	CovNets+DBN
Bar Chart	75.6%	45.6%	74.2%
Flow Chart	88.3%	56.8%	91.3%
Line Chart	71.2%	22.3%	67.9%
Scatter Chart	69.8%	44.5%	84.2%
Pie Chart	58.1%	50.1%	59.4%
Ave. Accuracy	72.6%	43.9%	75.4%

Table 3.3: Comparing Results of Proposed Framework from (Liu et al. [6])

4 Approach

As seen in the previous chapter, different approaches and methods have been described for chart image recognition. Due to the complexity of this task, a CovNet approach is employed for our work. Reasons for selecting this approach include; automatic capturing, learning and extraction of features; parameter sharing which allows the model to learn a single set of weights once, rather than a different set of weights every time [22] and more importantly a performance accuracy nearing human capability. In spite of all the advantages of CovNets, they are very computationally expensive to apply on high-resolution images. Fortunately, current GPU capabilities with the help of good optimization techniques can handle this issue [3].

The specific contributions of this thesis are as follows: to create a dataset of chart images (scatter-plot, bar-chart, line-chart and box-plot) using real-world data. This chart image dataset tries to capture all variety of structure in the various chart images to be handled. The scripts, libraries, and parameters used in creating the chart images are readily available for the recreation of the dataset. The second contribution of this thesis involves training a CovNet model for recognizing these types of chart images.

The rest of this chapter describes into details how the dataset was created, it also describes the CovNet architecture that was used for the classification and evaluation tasks.

4.1 Dataset

The following paragraph is inspired by a blog written on what to look for in training data [23]. The saying 'Garbage in Garbage out' is a valid statement when it comes to creating a dataset for machine learning. The machine learning technique will learn from whatever data fed to it. So if a dataset of good quality is fed into the algorithm, then the model created will also be of good quality. The dataset creation stage is therefore a very important stage. In most approaches that worked on chart images classification, the dataset used consisted of chart images downloaded from Google and a few others were ob-

tained from extracting chart images from pdf's. This approach we believe is limited since, we don't know the programming languages, the libraries used and parameters used in creating these charts. This missing information is very relevant since it tells us how diverse our resultant dataset is. For example, how are we sure that all the charts that were downloaded from Google, were not only created in Python or Java?, and in such a case how well will a new chart created with Matlab or R be classified. For this reason the dataset comprised of image charts created with different libraries and with dissimilar programming languages. Some random chart images from other works were manually labeled and added to our created dataset. In the next sections, the various datasets and the languages used in plotting are described. To make the dataset as diverse as possible, charts created in each programming language used a different set of CSV files.

4.1.1 Dataset for Matlab

The Data used for creating the plots in Matlab were randomly chosen from Project Dataset [24], a free CSV data repository, DatPlot [25] and Plotly CSV repository in github [26]. The datasets are multidimensional and compiled from normal day to day activities like dating, what makes people happy etc, and objects like cameras and cars. On the average the datasets used contain about 500 instance and 5 different columns. The biggest dataset is called Speed dating data. It is made up of over 8,000 observations of answers to survey questions about how people rate themselves and how they rate others on several dimensions. The smallest dataset used has 33 instances and 12 columns. It contains information about cars. The number of gears and speed, just to name a few attributes.

4.1.2 Dataset for R

For the plots in R, 13 random CSV files were downloaded from an archive of datasets distributed with R called Rdatasets [27]. Rdatasets is a collection of dataset distributed with R. On the average there are 80 instances and 5 columns in each dataset. The biggest CSV file is the Australian athletes dataset. It's made of 203 instances and 14 columns and contains attributes like sex,height,weight and sports. The smallest dataset is the Canadian Women's Labour-Force Participation. This dataset has 30 rows and 7 columns. It contains information like average wages of women, percent of adult women in the workforce etc.

4.1.3 Dataset for Python

The data used for creating the plots in Python were 15 randomly selected csv files also from Rdatasets [27]. The biggest dataset among the 15 is the Monoclonal gammopathy data, it contains natural history patients with monoclonal gammopathy of undetermined significance. The dataset is made up of 1384 observations with 10 columns, it has attributes like age, sex, time of death and last contact in months. On the average each dataset contains about 200 instances and 7 columns of multi-dimensional data. The smallest dataset however contains only 33 instances with 11 columns and is called the Nuclear Power Station Construction Data. The data relate to the construction of 32 light water reactor (LWR) plants constructed in the U.S.A in the late 1960's and early 1970's.

4.1.4 Dataset for Java

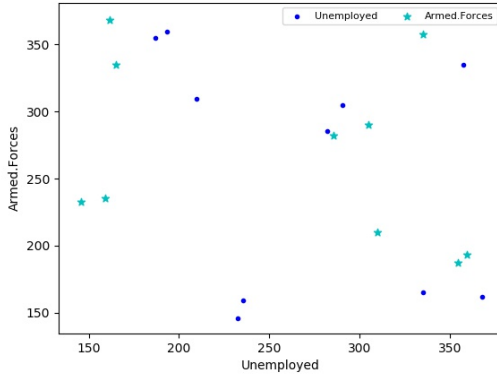
For the plots created in java, I used the dataset made available by Plotly [26], a github repository of CSV datasets used in the Plotly API examples. 14 random CSV files were downloaded, the biggest file has 1002 instances and 9 columns, and on the average each file contains about 100 instances and 9 columns. The smallest file however is made of 33 instances and 12 columns called the mtcars file. It contains information about a variety of different car models like the number of gears, speed etc. The table 4.1 contains the names of all CSV files that were used in the different languages with the different plotting programs.

Datasets			
Python	Matlab	R	Java
3d_line_sample_data.csv	Camera.csv	ais.csv	3d-line-plot.csv
LightForwardFlapStall.csv	Cars.csv	Angell.csv	3d-scatter.csv
line_3d_dataset.csv	speedDating.csv	Baumann.csv	2011_flight_paths.csv
longley.csv	Cereal.csv	Bfox.csv	2011_us_exports.csv
loti.csv	happiness.csv	cane.csv	auto-mpg.csv
lung.csv	TestData1.csv	carprice.csv	candlestick_dataset.csv
nuclear.csv	TestData2.csv	Chirot.csv	finance-charts-
timeseries.csv	mpg.csv	Davis.csv	apple.csv
USJudgeRatings	okcupid-	Ericksen.csv	globe_contours.csv
WVSCulturalMap.csv	religion.csv	Florida.csv	hobbs-pearson-
wind_rose.csv	spectral.csv	Highway1.csv	trials.csv
volcano.csv	stockdata.csv	Pottery.csv	motor_trend_tests.csv
uspop2.csvm	subplots.csv	Prestige.csv	nz_weather.csv
tips		salinity.csv	volcano.csv
		urine.csv	iris.csv
			mtcars.csv

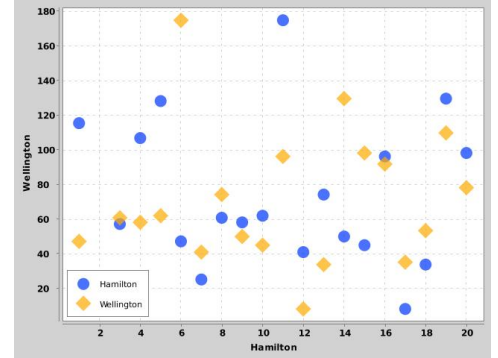
Table 4.1: Names of datasets used in each plotting program

4.2 Creating Plots

The motivation for creating a variety of plots to capture all type of plots used in scientific papers was acquired by inspecting the datasets of Architecture proposal for data extraction of chart images using CovNet paper [21] and Vizmetrics: Analyzing visual information in the scientific literature [28]. Scripts in various languages were written to handle the plotting and labeling process automatically. All datasets for a particular plot (example scatter plot for python) are put into one folder. The scripts reads each CSV file column by column while creating the plots. Table 4.2 describes how the plots where created in each language. The type column describes the different variety of a particular plot, for example bar charts can be of type stacked, grouped, vertical and horizontal bar charts, also scatter plots types can be a scatter plot consisting of one type of marker, one scatter plot with multiple markers and finally a scatter plot with a line showing the correlation between the plots. Figure 4.2 shows two different types of bar charts. Figure 4.2a is a stacked bar chart and Figure 4.2b a normal vertical bar chart. The Library column shows the different plotting libraries used, the parameter column describes parameters that were changed and finally the number of plots created were also added. The images below the tables are sample images that exist in our dataset of created plots for each language.



(a) Matplotlib scatter plot with star and circular markers



(b) Java scatter plot with circular and diamond markers

Figure 4.1: Scatter plots

4.2.1 Random Charts

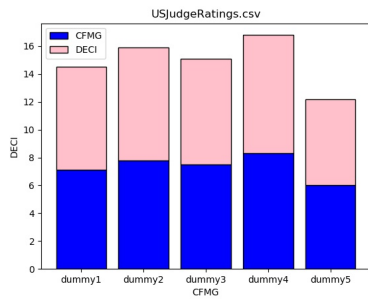
In order to have the most diverse dataset, so as to achieve a robust model, we decided to pick some random chart images from Junior et al. [21] and Po-shen et al. [29]. The line-charts, bar-charts and, scatter-plots were already labeled

Table 4.2: Overview of the varied parameters and libraries used for creating plots in the different plotting programs

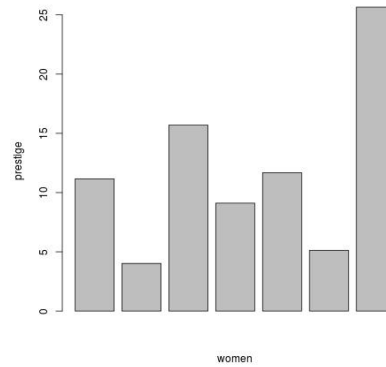
Scatter-plot	Language	Library	Parameters
	Python	Matplotlib v2.1.2 Plotly v2.5.1 Seaborn v0.8.1	MarkerStyle ['o', '*', '.', '+', 'x']
	MATLAB	Default Plotly	MarkerStyle ['o', '*', '+', 'x', 's']
	R	Plotly Lattice Ggplot2	MarkerStyle ['o', '*', '+', 'x', 's']
	JAVA	XChart 3.5.1 jfreechart 1.0.1	MarkerSize (15 -18)
Bar-charts	Python	Matplotlib v2.1.2 Plotly v2.5.1 Seaborn v0.8.1	
	MATLAB	Default	Width of bar(14-16)
	R	Default,Plotly R Library ggplot2	space (0-3)
	JAVA	XChart 3.5.1 jfreechart:1.0.192 javafx.scene	PlotOrientation (vertical or horizontal) with error bars
Line-chart	Python	Matplotlib v2.1.2 Plotly v2.5.1 Seaborn v0.8.1	Linestyle ['-', '--', '-.', ':']
	MATLAB	Default Plotly	MarkerStyle ['o', '*', '.', '+', 'x', 's'] markersize [8-10]
	R	Default,Plotly R Library ggplot2	
	JAVA	XChart 3.5.1 javafx JFreeChart	MarkerSize (12-16)
Box-Plots	Python	Matplotlib v2.1.2 Plotly v2.5.1 Seaborn v0.8.1	
	MATLAB	Default	
	R	Default,Plotly R Library ggplot2	
	JAVA	XChart 3.5.1 Jfree	LegendPosition (topleft,topright)

Table 4.3: Overview of the number of plot and different varieties plots in the different plotting programs

SCATTER PLOTS		
Language	Number of plots	Type
Python	1020	Unique markers, With legends, multiple markers regplot
MATLAB	1044	
R	1644	
JAVA	1644	
BAR CHARTS		
Language	Number of plots	Types(bar)
Python	1018	Horizontal and Vertical, Stacked, Grouped bar charts Error bars
MATLAB	1000	
R	1144	
JAVA	1144	
LINE CHARTS		
Language	Number of plots	Types(Line with)
Python	1000	Markers, Multiple Lines
MATLAB	1000	
R	1644	
JAVA	1644	
Box Plots		
Language	Number of plots	Types(Box with)
Python	1000	Notches, Multiple Boxes
MATLAB	1000	
R	1644	
JAVA	1644	

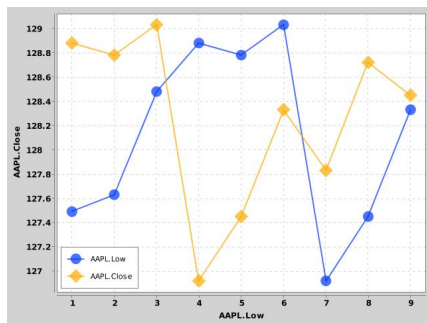


(a) Matlab stacked bar chart (bar width 16)

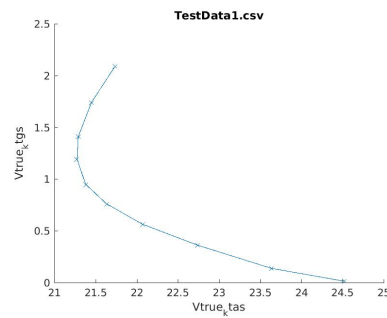


(b) R horizontal bar chart (bar width 16)

Figure 4.2: Example Bar Charts

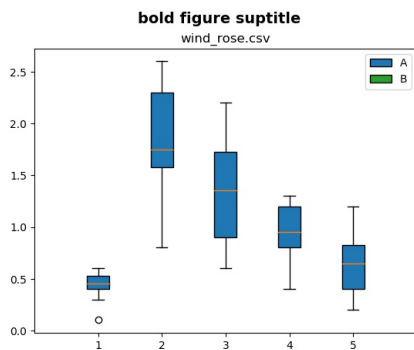


(a) Java line chart with diamond and circular markers

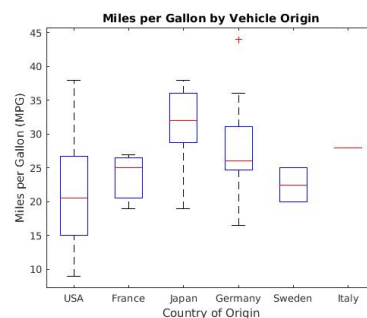


(b) simple Matlab line chart with Asterisk marker

Figure 4.3: Example Line Charts



(a) Vertical multiple boxplots in python



(b) Vertical multiple boxplots in Matlab

Figure 4.4: Example Box Plots

and were taken from Junior et al. work and the box-plots which we had to label, were taken from Po-shen et al. work since, the prior did not contain box-plots. The randomly picked chart images, were then added to the already created dataset for training and testing. Figure 4.5 shows sample images from the already created chart images from the paper of Po-shen et al. and Junior et al.

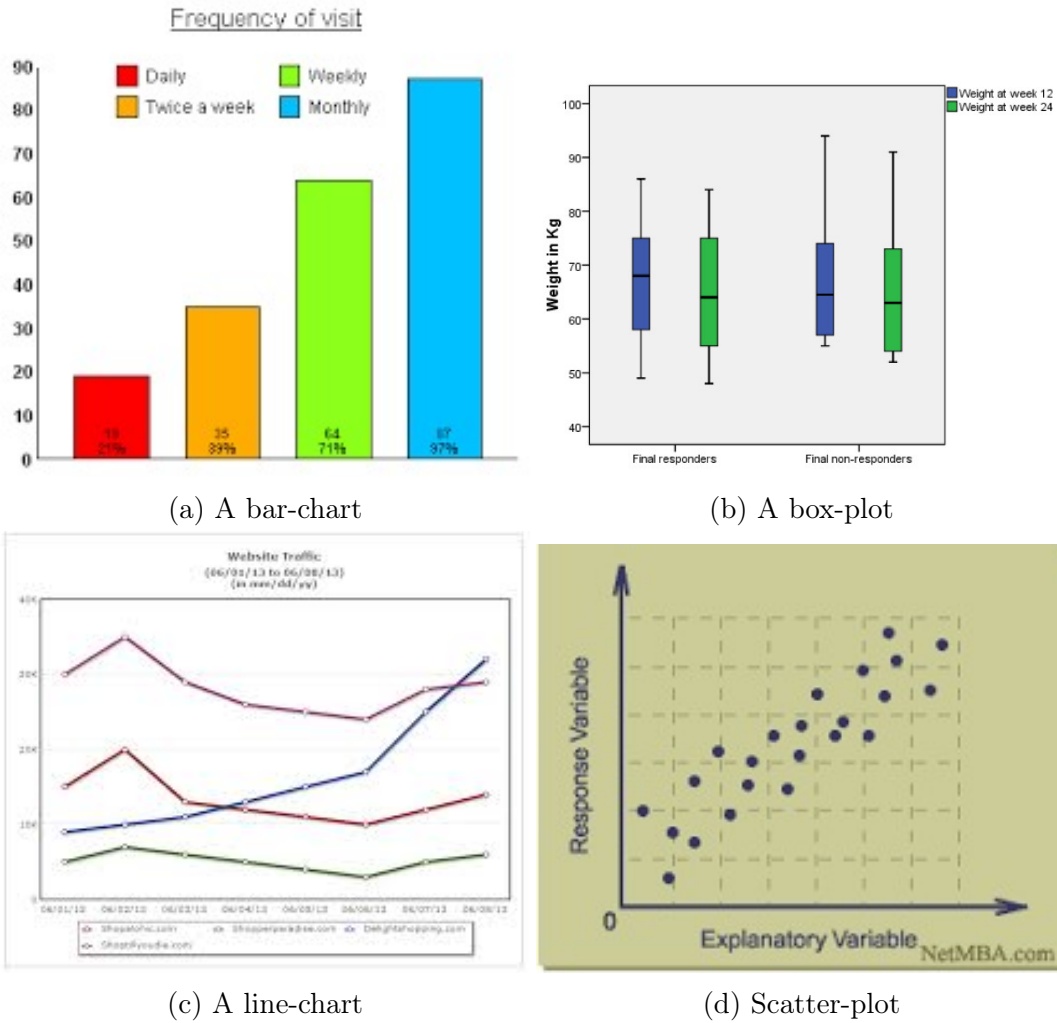


Figure 4.5: Example Chart images from Po-shen et al. and Junior et al.

4.2.2 Training, Validation and Test set

Figure 4.6 summaries how our dataset was split into various activities. In machine learning, a model is an artifact created during the training phase, this model can be likened to a function that maps specific features to their respective labels, the model is trained with a portion of the dataset called the train set. As the model is trained, the validation set is used to decide and pick which metric out of hyper-parameters, early stopping and architecture considerations yields the best performance. This helps to adjust and optimize

the model [30]. Finally, the test-set in machine learning, is the other portion of the dataset which was not used during the training phase. The test-set checks how well the trained model performs on unseen data by giving an unbiased assessment of the model. For this work, the dataset is divided into train, validation and test set. The validation set is not shown in the figure since it forms part of the train data.

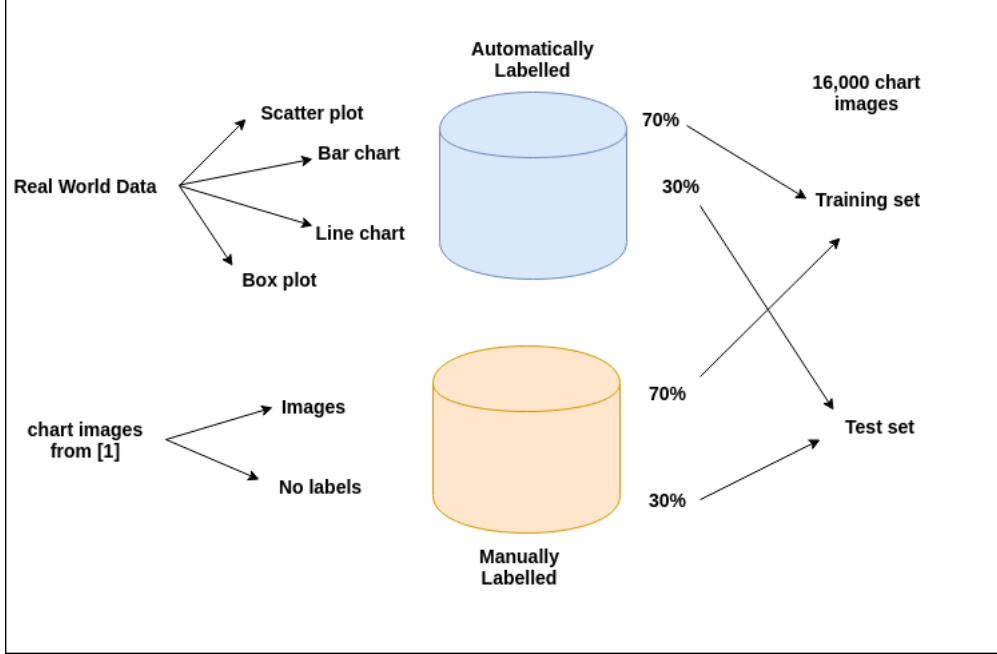


Figure 4.6: Steps involved in building the Classification model

4.3 The Architecture

As mentioned in previous sections, our model is trained with a CovNet. This architecture is inspired by the famous AlexNet CovNet architecture [3]. The AlexNet architecture was used in the famous ImageNet LSVRC-2010 contest to classify 1.2 million into 1000 classes, the results obtained on the test-set had a top-5 error rate of 17.0%, which was better than the previous state of the art. Our CovNet architecture just like the AlexNet architecture contains 8 learned layers, these 8 layers are made up of five Convolutional layers and 3 fully connected layers. A Relu is used after some Convolutional layers and fully connected layers. Another parameter employed in our architecture is a dropout, a dropout is applied in the first and second fully connected layers of our network and finally, max pooling is applied in some of the layers. A summary of our CovNet architecture used to train our model is shown in Figure 4.7. The preprocessing steps, the parameters used, and why they were employed are explained into details in the rest of this chapter.

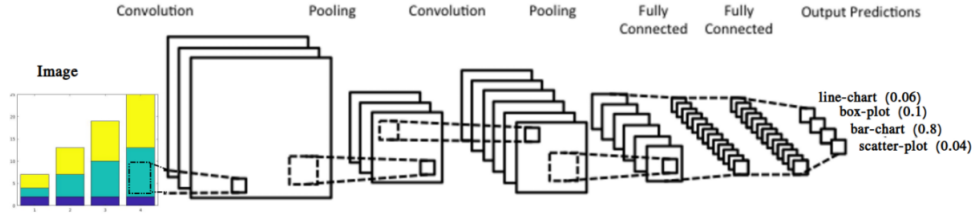


Figure 4.7: Architecture of CovNet used to train our model

Convolutional Layer

As mentioned above 5 of this layers are used in our architecture, and these are responsible for extracting the features of the chart images. Our input image size used is $224 \times 224 \times 3$ since its an RGB image, this is fed into our first Convolutional layer, this filters the image with a stride of 4 pixels. The second Convolutional layer takes as its input the output of the first Convolutional layer, between this two layers max-pooling is performed to reduce the size of the image. The third and fourth Convolutional layers, unlike the previous two have no max-pooling applied between them. The final Convolutional layer then takes as input the output of the previous layer and performs a max-pooling.

4.3.1 Dropout

A dropout was used in the first and second Fully connected layers in our architecture. We set dropout probability to 0.5 in our network, this means nodes are dropped out with a probability of 1-05 or or maintained at a probability of 0.5.

4.3.2 Fully connected layers

The last set of layers used are the Fully connected layers. The first layer takes as input the 1D feature vector generated by a flatten layer in our network. The flatten layer converts all our 2D arrays from our previous Convolutional layers into a 1D resultant vector. The output of this layer is then fed as input into the second fully connected layer. The output of the second layer is then fed into the last fully connected layer. The last layer in this set of 3 layers uses the softmax activation to give our output prediction. The output is a separate probability of our four classes and the choice with the highest probability is our chosen prediction. We used a softmax activation as opposed to other activation functions because, it handles low invigoration in images (say blurry charts) of your neural net with rather uniform distribution and to high invigoration in images (say large numbers, like sharp images) with prediction probabilities between 1 and 0.

Operation	Filter	Depth	Stride
Conv1 + Relu	11 * 11	96	4
Max Pooling	3 * 3		2
Conv2 + Relu	5 * 5	256	
Max Pooling	3 * 3		2
Conv3 + Relu	3 * 3	384	
Conv4 + Relu	3 * 3	384	
Conv5 + Relu	3 * 3	256	
Max Pooling	3 * 3		2
Dropout(rate 0.5)			
FC6 + Relu			
FC7 + Relu			
FC8 + Relu			

Table 4.4: The following table shows different layers, parameters and computation units used.

5 Experiments

6 Summary

7 Bibliography

- [1] Matange, S. Box plot legend. <https://blogs.sas.com/content/graphicallyspeaking/2017/04/13/box-plot-legend/>, 2018. [Graphically Speaking; accessed 18-Jun-2018].
- [2] karpathy. Cs231n convolutional neural networks for visual recognition. <http://cs231n.github.io/linear-classify/#softmax>, 2018. [Online; accessed 11-Jun-2018].
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [4] Nitish Srivastava, 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(1):1929–1958, 2014.
- [5] V Karthikeyani and S Nagarajan. Machine learning classification algorithms to recognize chart types in portable document format (pdf) files. *International Journal of Computer Applications*, 39(2), 2012.
- [6] Xiao Liu, Binbin Tang, Zhenyang Wang, Xianghua Xu, Shiliang Pu, Dapeng Tao, and Mingli Song. Chart classification by combining deep convolutional networks and deep belief networks. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 801–805. IEEE, 2015.
- [7] Few, S. Visual business intelligence – abela’s folly – a thought confuser. <https://www.perceptualedge.com/blog/?p=2080>, 2018. [Perceptualedge.com.; accessed 17-Jun-2018].
- [8] Thermopylae Sciences + Technology. Humans process visual data better. <http://www.t-sciences.com/news/humans-process-visual-data-better>, 2018. [Thermopylae Sciences + Technology. (2018).; accessed 18-Jun-2018].

- [9] Manolis Savva, 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*, pages 393–402. ACM, 2011.
- [10] V Shiv Naga Prasad, Behjat Siddiquie, Jennifer Golbeck, and Larry S Davis. Classifying computer generated charts. In *Content-Based Multimedia Indexing, 2007. CBMI'07. International Workshop on*, pages 85–92. IEEE, 2007.
- [11] Jihen Amara, Pawandeep Kaur, Michael Owonibi, and Bassem Bouaziz. Convolutional neural network based chart image classification. 2017.
- [12] Richard Boyle, Bahram Parvin, Darko Koracin, Nikos Paragios, and Syeda-Mahmood Tanveer. *Advances in visual computing*. Springer, 2007.
- [13] Yan Ping Zhou and Chew Lim Tan. Bar charts recognition using hough based syntactic segmentation. In *International Conference on Theory and Application of Diagrams*, pages 494–497. Springer, 2000.
- [14] Mingyan Shao and Robert P Futrelle. Recognition and classification of figures in pdf documents. In *International Workshop on Graphics Recognition*, pages 231–242. Springer, 2005.
- [15] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 13–23. Springer, 2000.
- [16] Cornelisse, D. An intuitive guide to convolutional neural networks. [online] freecodecamp. <https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050>, 2018. [Online; accessed 12-Jun-2018].
- [17] Nikhil B. Image data pre-processing for neural networks – becoming human: Artificial intelligence magazine. <https://becominghuman.ai/image-data-pre-processing-for-neural-networks-498289068258>, 2018. [Online; accessed 13-Jun-2018].
- [18] Balwally, A. What is the role of rectified li. <https://www.quora.com/What-is-the-role-of-rectified-linear-ReLU-activation-function-in-CNN>, 2018. [Online; accessed 9-Jun-2018].

- [19] Rodriguez, J. Convolutional neural networks for the rest of us part iii: Benefits and motivation. <https://medium.com/@jrodthoughts/convolutional-neural-networks-for-the-rest-of-us-part-iii-benefits-and-motiv> 2018. [Online; accessed 10-Jun-2018].
- [20] ujjwalkarn. An intuitive explanation of convolutional neural networks – the data science blog. <https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/>, August 2016. (Accessed on 05/16/2018).
- [21] Paulo Roberto Silva Chagas Junior, Alexandre Abreu De Freitas, Rafael Daisuke Akiyama, Brunelli Pinto Miranda, Tiago Davi Oliveira De Araújo, Carlos Gustavo Resque Dos Santos, Bianchi Serique Meiguins, and Jefferson Magalhães De Moraes. Architecture proposal for data extraction of chart images using convolutional neural network. In *Information Visualisation (IV), 2017 21st International Conference*, pages 318–323. IEEE, 2017.
- [22] Rodriguez, J. Convolutional neural networks for the rest of us part iii: Benefits and motivation. <https://medium.com/@jrodthoughts/convolutional-neural-networks-for-the-rest-of-us-part-iii-benefits-and-motiv> 2011. [Online; accessed 5-Jun-2018].
- [23] Editor. what-to-look-for-in-training-dataset. <https://medium.com/@jrodthoughts/convolutional-neural-networks-for-the-rest-of-us-part-iii-bene> 2018. [Online; accessed 5-Jun-2018].
- [24] James Eagan. Project datasets. <https://perso.telecom-paristech.fr/eagan/class/igr204/datasets>. [Online; accessed 20-April-2018].
- [25] Michael Vogt. Datplot. <https://vincentarelbundock.github.io/Rdatasets/datasets.html>, 2011. [Online; accessed 20-April-2018].
- [26] plotly/datasets. Latex — Wikipedia, the free encyclopedia. <https://github.com/plotly/datasets>, 2011. [Online; accessed 20-April-2018].
- [27] vincentarel bundock. Rdatasets. <https://vincentarelbundock.github.io/Rdatasets/datasets.html>, 2011. [Online; accessed 20-April-2018].
- [28] Po-shen Lee, Jevin D West, and Bill Howe. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, 4(1):117–129, 2018.

- [29] Po-shen Lee, Jevin D West, and Bill Howe. Viziometrix: A platform for analyzing the visual information in big scholarly data. In *Proceedings of the 25th international conference companion on world wide web*, pages 413–418. International World Wide Web Conferences Steering Committee, 2016.
- [30] Akshay Balwally,. What is the difference between validation set and test set? <https://www.quora.com/What-is-the-difference-between-validation-set-and-test-set>, 2016. [Online; accessed 7-Jun-2018].