FRI Data Science Project Competition 2020
Interim report

How good is my plot?

Maruša Oražem, Urša Zrimšek, Andrej Hafner

Abstract

The goal of this project is to build a model that will evaluate the quality of a given plot. We constructed a dataset of different plot types, from which we will gather the opinions on quality through a crowdsourcing platform. From these we plan to find the correspondences between visual properties and plot quality.

Keywords
visualizations, plots, quality perception

Advisors: Assoc. Prof. dr. Erik Štrumbelj , Assoc. Prof. dr. Tomaž Curk

Introduction

Statistical plots are one of the basic tools, which a scientist can use to convey information. They are used and abused in many scientific works, and are unfortunately often not of a good quality. Therefore, the goal of our project is to develop a model that can find out what are the properties of a high-quality plot? The first thing that comes to mind is how well the plot presents the information in the data.

Different types of plots are suitable for different kinds of data. A line plot is great for presenting time-series data, while it fails on categorical data. On the other hand, box plots can be great for categorical data, but inconvenient for continuous data. Since evaluating the quality of a plot on such a level requires a deeper understanding of the context, we focus mainly on the visual aspect. This enables us to utilize deep learning to find a relationship between the visual properties of the graph and its quality.

What makes a good plot?

The biggest question we have to answer is, what makes a good plot? To get an overview on that, let's summarize the parts of the article *Testing statistical Charts: What makes a Good Graph?* [1], that we can (at least implicitly) use in our approach of evaluating the plot's quality.

The main goal of statistical plots is to effectively and accurately present the data. Any additional enhancement on the plot should contribute to that aim. Here are some things that are considered to be current best practices. Let's first list some of the cognitive principles:

- **Proximity**. This principle tells us, that we need to put things we want to compare closer together, just the opposite for the less important things.
- **Similarity**. This principle tells us to group together things that somehow belong to the same group. We could do that with the same color or same shape.
- **Common region**. This principle suggests that elements that are contained in a common region, belong together.
- **Working memory**. This tells us about the limitation of our working memory. In general, we can not process more than $7(\pm 2)$ items.
- **Change blindness**. This principle talks about the fact, that it is difficult to detect changes on a similar object.
- **Ease of comparison**. Some things are easier to compare than others. If things can be compared in a way that is higher in the following hierarchy, it should be done that way:
  1. position (common scale),
  2. position (nonaligned scale),
  3. length, direction, angle, slope,
  4. area,
  5. volume, density, curvature,
  6. shading, color saturation, color hue,
  7. discriminable shape,
  8. indiscriminable shape.

It was shown that aesthetics used in plots can change how the data is read. A very important thing we need to consider is the colors. Color can change everything. Like we already stated, elements of the same group should have the same color. But what is that color? Colors should be chosen to best match the data and plot type. If we are trying to show magnitude, we should use a univariate color scheme. On the other hand, if we want to show the data that differ in sign, we should use a double-ended color scale. If possible, we should keep the number of hues to a minimum, we should transit through neutral colors, such as white when utilizing a gradient. Because color blindness is common in society, we should use color-blind accessible colors. These properties are a generalization of what is common in our perception, each person still has their own special preferences. Since we can't directly grade a plot based on these categories, especially if we don't know the context, we will rather utilize a crowdsourcing platform to acquire relevant data.

Getting the data

The first task was to get the statistical plots to work with. We have decided to get them from different resources. We have chosen 7 different plot types – box plot, line plot, bar plot, histogram, pie chart, scatter plot and violin plot. Then we have scrapped these types of plots from Google, Bing, and DuckDuckGo image search results. We got approximately 400 images for each class. Next, we have downloaded bachelor's, master's and doctoral thesis from *Repository of the University of Ljubljana* [2] and extracted images from there. Finally, we have also downloaded images from the *PLOS One* journal [3].

Classification model

Now we gathered datasets of different images. The images scraped from search engines are already categorized into correct plot classes. We filtered them and deleted bad images by hand, and then used them to train a convolutional neural network (CNN) to help us with the classification of other images. The majority of images extracted from PDF-s were not plots, so we added the eight class called `not_plot`. For its training set, we used 400 randomly selected images, extracted from the FRI bachelor's theses, from which we deleted any image that represented a plot.

For classification, we used a residual neural network, more precisely ResNet101, which has 101 layers. Since our train dataset is not large, we needed to take a pretrained network, that was trained on more than a million images of 1000 classes from ImageNet. Pretrained CNN is already good at recognizing representative shapes, we just need to fine-tune it to be able to separate between the classes we are interested in. We deleted the last layer and replaced it with a fully connected layer with 8 neurons, each representing one of our classes. Then we used Google Colab to do 50 epochs of training on our dataset, using Adam optimizer together with cross-entropy loss. We validated it in each epoch and saved the model that had the lowest loss. We tried different batch shapes and learning rates, and reached best validation accuracy of 95% and cross-entropy loss of 0.16.

After the first iteration of training, we classified the first 2000 images from FRI bachelor's theses and sorted them into 3 groups – those for which the model returned more than 99% probability for the predicted class, those for which it returned more than 90% and those for which it returned less. We then handpicked all wrongly classified images from there, and also those correctly classified, that the model wasn't sure about, and added them to the training set. On figure 1 we can see some examples.

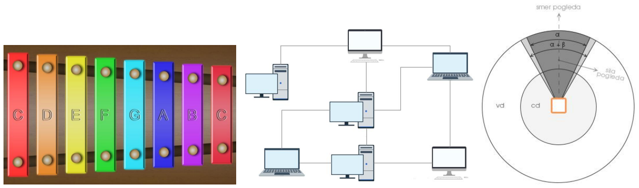

Figure 1. Images that were wrongly classified as plots. From left to right we can see images classified as bar plot, box plot and pie chart. None of them is a plot, but if we think in the context of CNNs, we can understand why the model made those mistakes.

In the second phase, when we trained the network again together with those harder examples, it reached 96% accuracy and a loss of 0.13 on validation data. We will use this model for filtering the extracted images, and will additionally fine tune for different sources, if needed.

Further work

Now we are trying to figure out, what are the attributes which define whether a plot is good or not. For determining the most popular attributes, we have decided to design some questionnaires and give it to a few people. On those questionnaires will be different plots. For each pair, an individual will decide which one is better and why. For instance *"left plot is better because it uses appropriate colors", "right one is better because the axis are labeled",* ... When we will get the results from several questionnaires, we will gather the results and figure out what are the most important attributes to consider when evaluating the quality of plots. Next, we will use Amazon's Mechanical Turk to evaluate our plots data set. Workers will be asked similar questions, but now on attributes we have gathered from questionnaires. The plan is to evaluate pairs of pictures, for each which one has better *attribute1, attribute2,* ... After we will have all the data, we are going to build a regression model that will predict how good the plot is and what are the values of its attributes – those should explain why the plot is good/bad.

References

[1] Heike Hofmann Susan Vanderplas, Dianne Cook. Testing Statistical Charts: What makes a good graph? *Annual Reviews*, 6, 2020.

[2] Repository of the University of Ljubljana. https://repozitorij.uni-lj.si/info/index.php/eng/. Accessed: 22-4-2021.

[3] PLOS ONE journal. https://journals.plos.org/plosone/. Accessed: 22-4-2021.