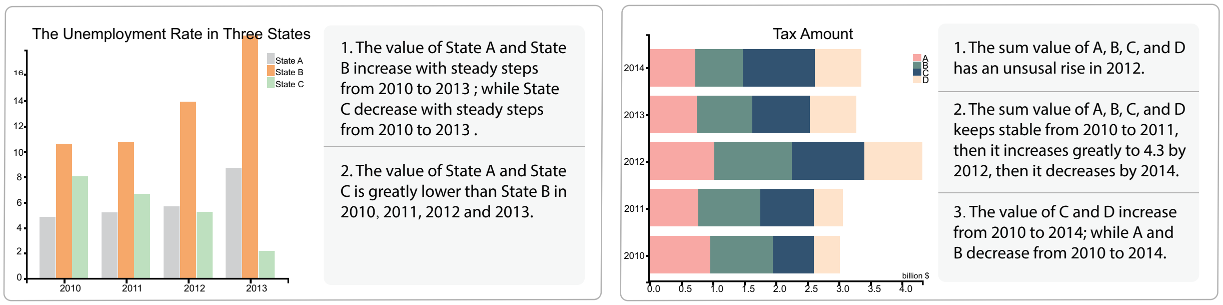
Automatic Caption Generation for Visualization Charts

Can Liu[1] Liwenhan Xie[1] Yun Han[1] Xiaoru Yuan[1,2,3*]

1) Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, China
2) National Engineering Laboratory for Big Data Analysis Technology and Application, Beijing, China
3) Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University, Beijing, China



1. The value of State A and State B increase with steady steps from 2010 to 2013 ; while State C decrease with steady steps from 2010 to 2013 .

2. The value of State A and State C is greatly lower than State B in 2010, 2011, 2012 and 2013.

1. The sum value of A, B, C, and D has an unsusal rise in 2012.

2. The sum value of A, B, C, and D keeps stable from 2010 to 2011, then it increases greatly to 4.3 by 2012, then it decreases by 2014.

3. The value of C and D increase from 2010 to 2014; while A and B decrease from 2010 to 2014.

Figure 1: A showcase of captions automatically generated based on data with certain features. The left one shows different trends of three states; The right one highlights a sudden rise in 2010.

Abstract

Captions act an important role in guiding people to interpret the chart and conveying messages from the designer. But it requires labor efforts to make a proper caption. In this paper, we propose a novel automatic approach to generate captions from visualization charts powered by deep learning. The model learns to recognize significant features of the chart, which are mainly represented by subsets of its visual elements. Through a carefully designed summary template, each subsets is converted into a descriptive sentence, i.e. data fact, and compose a complete caption for the chart.

1 Introduction

Information graphics such as bar charts and line plots are important components of daily-life usage to demonstrate certain facts like trends, extremes and so on. Typical visualizations in an article are composed of graphs along with a caption emphasizing crucial information as well as pointing out intended messages. A proper caption is the very key for audience to interpret the charts fast. However, it remains a mechanical labor work for chart designers to make captions, manually converting the chart features into descriptive sentences. Currently, automatic caption generation for charts is still in its infancy. Most works in the literature are based on statistical functions and heuristics, which introduce a dilemma to pick the number of facts : an excessive list may lead to cognitive overload, while too little facts may cause the lost of critical information.

In this paper, we propose an approach with deep learning that enables automatic extraction for crucial facts from the charts. Rather than providing overmuch details, captions generated through our method gives the main information in a small number of facts. The deep-learning model learns the way people extract features from the chart and outputs several subsets of visual elements, which represent facts separately and are converted into sentences using summary templates. To validate the effectiveness of our model, we experiment with the bar chart of various forms and obtain promising results.

2 Method Overview

Our approach takes an SVG-based chart as input and outputs a caption composed of sentences, each of which corresponds to a crucial fact of the input chart. The workflow could be divided into three steps: 1). **Data parsing**: Parse graphical elements with attributes from the chart to get the composed elements and then extract corresponding data. 2). **Fact extraction**: Extract crucial facts worth to describe from those elements powered by 1-D residual network with GAN. 3). **Template-based caption generation**: Generate sentences that make up the caption for each fact based on templates.

2.1 Data Parsing

We present a statistic-based method to find the mapping between element attributes and the origin data. Firstly, we extract the element groups from the SVG chart. Next, we distinguish the axis and legend from the table. Leveraging the ticks and text from the axis and rect - text pairs from the legend, we derive the quantitative or category mapping relationship from those components. Further, we extract the data from graphical elements based on such mappings, after removing occasional irrelevant graphical elements with heuristics.

2.2 Fact Extraction

Following a common scheme to use tensor in machine learning, we present input **data elements** and the output **facts** in the form of a tensor. Formally, we define the **data element tensor**. The data element can be presented as $V = \{e_{1m}, e_{2m}, ...e_{nm}\}$ where $e_{im} = \{attrs_{visual}\}$, $attrs_{visual}$ stands for visual attributes, i.e., position, color, etc. With the parsed corresponding data attributes, data elements in tensor form can be further presented as $V = \{e_{1m}, e_{2m}, ...e_{nm}\}$ where $e_{im} = \{attrs_{visual}, attrs_{data}\}$, each element here coupled visual attributes with data attributes.

The data attributes have different types, e.g., quantitative, ordinal, and categorical types. We develop various encodings for each types.

Fact tensor presentation: For the model to present the information, it's more convenient to present the fact in a tensor format.

[*] Email: {can.liu, xieliwenhan, yunhan, xiaoru.yuan}@pku.edu.cn; Xiaoru Yuan is the corresponding author.

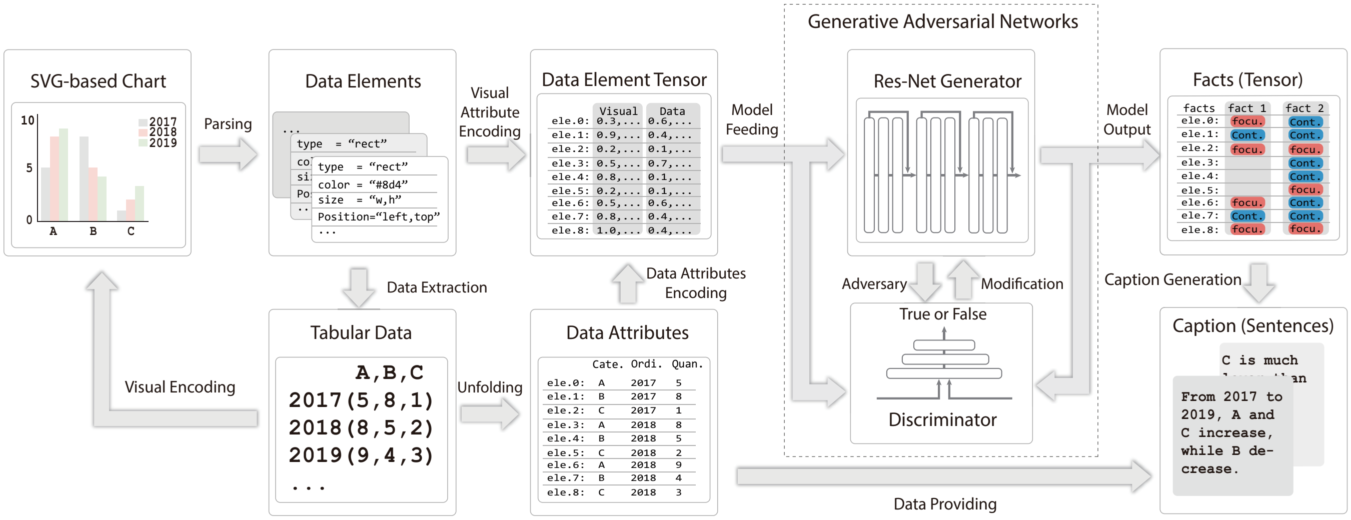


Figure 2: The workflow of our approach to generate captions for charts. An input SVG-based chart is parsed into data elements with a form of tensor, which are later fed to the GAN scheme embedded with Res-Net generator and then outputs an **fact tensor** indicating fact type and fact elements. Using a summary template, we translate such a tensor into a caption in natural language.

A fact tensor $V_{fact} = \{v_1, v_2, \ldots, v_n\}$ is defined as a 3-channel tensor, where: $v_i = [0,0,1],\ if\ e_i \in F$; $v_i = [0,1,0],\ if\ e_i \in C$; and $v_i = [1,0,0],\ if\ e_i \notin (F \cup C)$.

Several facts can be combined by stacking their channels. In Figure 3, the right part shows the facts tensor with three facts which are stacked by the channels.

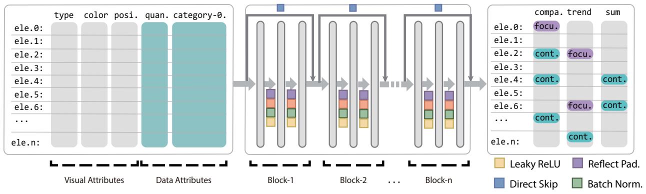


Figure 3: The generator of the fact extraction part. Each component in the block has a 1D-convolutional layer, a Batch normalization layer and a ReLu layer.

The position of a fact tensor is called a **slot**, which encodes the fact type. For instance, Figure 3 shows that the fact in the first slot is in the type of comparison. The type determines what caption generation template is to be used.

2.2.1 1D-Residual Convolutional Network

We implemented a generator based on the 1D-residual convolutional network [1]. This 1D-dimension stands for the elements dimension. For the input, it's apparent that the channels represent the attributes of the major element. Figure 3 shows the generator. Rather than going through the networks layer by layer, the residual networks have a direct skip connection across three stacked nonlinear components.

2.2.2 Generative Adversarial Networks (GAN) Scheme

After experiments, we couple the GAN scheme with 1-D residual network to achieve the best performance. In the discriminator, the input is a tensor concatenated from the data elements tensor and the facts tensor. There are two kind of pair as training input; (1) pairs of data elements and generated facts tensor; and (2) pairs of data elements and ground truth facts. The pair with generated facts will be labeled as **false** while with ground truth facts labeled **true**.

2.3 Template-based Caption Generation

Information charts mainly show the following kinds of insights [2]. **Aggregation**: reduce the data item by statistic values like maximum, minimum, average or sum. **Trend**: the quantitative value changes by the ordinal attributes. **Comparison**: a focus on difference among items, item groups and trends. Under each class, or cross classes are sub-class or cross-class fact types, e.g., compare trends, compare aggregations, etc. According to the facts types above, we make a summary and produce templates to generate chart description. In this way, we then replace the objects and relations in the sentence and obtain a new sentence for this fact.

3 Results

Figure 1 showcases our results for charts beholding certain features and demonstrates the effectiveness of our proposed approach.

4 Conclusion & Discussion

In this work, we employ deep neural networks to extract crucial elements from visualization charts and achieve the task of automatic caption generation. Although the current implementation is limited to SVG-based charts, the general pipeline could be extended to other formats. Potential application scenarios of our approach include information retrieval from web-based charts, fast caption generation for data news. Further, combined with text-to-speech techniques, our method could be used to promote data accessibility for people with visual impairment.

Acknowledgments

This work is supported by NSFC No. 61872013 and the National Key Research and Development Program of China (2016QY02D0304). This work is also supported by PKU-Qihoo Joint Data Visual Analytics Research Center.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016.

[2] T. Munzner. *Visualization analysis and design*. AK Peters/CRC Press, 2014.