AutoCaption: An Approach to Generate Natural Language Description from Visualization Automatically

Can Liu\textsuperscript{1,*} Liwenhan Xie\textsuperscript{1}† Yun Han\textsuperscript{1}‡ Datong Wei\textsuperscript{1}§ Xiaoru Yuan\textsuperscript{1,2,¶}

\textsuperscript{1}Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, China
\textsuperscript{2}National Engineering Laboratory for Big Data Analysis and Application, Peking University, Beijing, China

\textbf{Abstract}

In this paper, we propose a novel approach to generate captions for visualization charts automatically. In the proposed method, visual marks and visual channels, together with the associated text information in the original charts, are first extracted and identified with a multilayer perceptron classifier. Meanwhile, data information can also be retrieved by parsing visual marks with extracted mapping relationships. Then a 1-D convolutional residual network is employed to analyze the relationship between visual elements, and recognize significant features of the visualization charts, with both data and visual information as input. In the final step, the full description of the visual charts can be generated through a template-based approach. The generated captions can effectively cover the main visual features of the visual charts and support major feature types in commons charts. We further demonstrate the effectiveness of our approach through several cases.

\section{1 Introduction}

Visualization designers create natural language caption (NLC) to interpret visualization for a better propagation among users. NLC is helpful in conveying crucial information on visualizations in data news and academic papers. However, the visualization community pays little attention to generate NLC automatically, which is helpful when lots of visualization do not have NLC. The automatic generation of NLC would be useful for data analysts like data news designer and stock analysts. The reason why there is no such algorithm to generate NLC directly is that natural language involves a higher level of human cognition, while traditional methods are hard to handle the complex task to summarize a visualization. Nowadays, the development of deep learning technology brings the possibility to deal with this problem. Deep learning methods have shown their ability in a similar task like image captioning \cite{16, 19}, which is similar to visualization captioning task. Automatic image captioning also brings significant benefits to image understanding and retrieval. Based on this, deep learning technology should have the potential to solve the automatic generation of NLC in visualization. However, to generate natural language from a visualization chart requires a long term exploration. As the first step of exploration, we present a machine learning-based pipeline to generate natural language caption automatically from a visualization.

Our approach supports charts in vector format (e.g., SVG format), which is a standard format in nowadays social media and online news feeds. Additionally, our approach can be extended to bitmap format charts leveraging Object Detection and OCR techniques. It is easy to recognize visual elements because they are regular shapes like rectangles and circles. The learning model detects crucial features of the chart, each of which is defined by feature type and grouped elements. A template-based method converts these features into natural sentences. Then we construct a caption for the chart leveraging the extracted weighted facts from the model.

To validate the effectiveness of this model, we experiment with bar charts, line charts, and scatter plots of various forms. From the result of validating dataset and tests on specific synthetic data, it is shown that the model can extract crucial features. We then deploy the pre-trained model onto a demo system and conduct a survey to people from different areas. The system supports accepting an uploaded chart. The system supports the exploration of each generated sentence in the chart with their corresponding elements highlighted in the chart display view.

In this work, the primary contributions are three-fold:

- We introduce a novel approach to generate captions for information charts with a deep learning-powered scheme, which learns the noteworthy features in accordance with human perception.
- We design a new model that is extensible to different charts with various suites of summary templates.
- Our prototype system demonstrates a potential usage of our approach in education and data overviews combining with practical annotations.

\section{2 Related Work}

Our work draws on prior researches in the area of chart interpretation, automation in visualization, and deep learning for caption generation in natural images.

It has long been the interest of the computational linguistics community to generate a text summary for an information graphic automatically, as this would promisingly improve the data accessibility of multimodal documents \cite{3}. Due to the challenge of a wide diversity in chart styles, most works focus on a certain type of chart modeling its context with rules and then exploring approaches to generate summaries, such as line charts \cite{18}, bar charts \cite{4}, and pie charts \cite{7}. In most works, rules are obtained through a series of perception cognition experiments \cite{6}. The influential work from Mittal et al. \cite{10} propose a method to describe the visual mapping in natural language, but they require users to give information about the mapping relationships. Al-Zaidy et al. \cite{1} use a semantic-graph representation for bar charts and then generate the summary by the “protoform” method \cite{20}. Bryan et al. \cite{2} use pre-defined scores for types of summaries to generate annotations for temporal charts. Hullman et al. also provides rule-based annotations for stock visualization \cite{9}. There is no previous approach to learn the chart features automatically. However, there is an increasing interest in automatic knowledge discovery from the database \cite{5, 15}. Additionally, many industrial products like Arria\textsuperscript{1}, Google Sheets\textsuperscript{2}, etc. employ Natural

\textsuperscript{1}Arria: https://www.arria.com/core-tech/
\textsuperscript{2}Google Sheets: https://www.google.com/sheets/about/
Language Generation (NLG) techniques to suggest such insights. In general, they use a suite of algorithms to find out statistically important results and generate natural language with a pre-defined template [12].

In addition to the efforts above, several works devoted to facilitating chart understanding with additive annotations, where automatic chart interpretation is also necessary. Srinivasan et al. [14] obtains the inferential data facts from a set of heuristics, and then suggest visualizations and embellishments to users after analyzing relationships between common descriptions, charts, and annotations.

3 Method

Two things are important to generate a caption of visualization, which are getting the content and presenting the content using natural language. We first have a step that parses the mapping relationship. For this step, there have been many previous works to parse a visualization into underlying data [11, 13]. Our main contribution is not revealing the underlying data, so we adopt existed method to handle this part [11]. Data information is recovered by parsing the visual information of data elements with the mapping relationship.

The next steps are to extract the feature content and generate the sentence. The process can be divided into two parts—Chart feature detection. We proposed a deep-learning model which takes the visual and data information as input and finally generate the chart feature. Each chart feature has a chart feature type and divides the elements into several groups. Sentence generation. We provide a default template for each type of chart feature. We use the output elements of the model to calculate the filling template components.

### 3.1 Item Definition

To describe the overall pipeline better, we first define the following concepts.

- **Data elements** is the elements in the charts that have the corresponding data item in the underlying data. Take the bar charts as an example. The bars are the data elements. The data elements can be presented as \( V = \{ e_1m, e_2m, ... , e_m \} \) where \( e_im = \{ \text{attr}_{visual} \} \), \( \text{attr}_{visual} \) is the visual information, i.e., position, color, opacity, etc.

- **Chart feature** is the content of a sentence. A chart feature has a feature type and several groups of data elements. The type corresponds to a certain template, while the related elements of a feature are composed of grouped elements; each group is a subset of data elements. Formally, \( \text{Feature} = \{ \{ G_i \}, \text{type} \} \), where \( G_i = \{ f_1, f_2, ..., f_m \} \) and \( C = \{ e_1, e_2, ..., e_k \} \), where \( 0 \leq m, k \leq n, G_i \in V_c \). Information charts mainly show features in the following types. Aggregation reduce the data item number by some statistic values such as maximum, minimum, average, or sum. Trend shows how quantitative value changes by the ordinal attributes. Comparison presents the difference, which often grabs the users’ attention. Distribution like the cluster and outliers are the patterns user often observed. Table 1 list the common features and the corresponding description template and example.

- **Description**: A description is a natural language presentation of a chart feature, which is generated from the chart type and relevant elements list through our summarized templates. The details of translating the chart feature into a description will be described in section 3.3.

- **Caption**: A caption is composed of several chart feature descriptions.

### 3.2 Chart Feature Detection

This part is to extract the chart feature from the chart. He et al. [8] introduces the residual network (Res-Nets) in the task of 2D-image recognition, which can well handle the deeper networks which well handled the problem of gradient vanishing. The residual networks have direct skip connections across several layers. Unlike the original Res-Nets with two-dimension layers that deal with 2D-image tasks, our model tackles the 1D-tensors. Residual CNN keeps the sizes of the tensor, which can handle charts with different number of elements. A desired underlying mapping that encourages the model to approximate is \( H(x) \). Now with the direct connection sent the input x directly to the output, the stacked components only need to approximate \( F(x) \), where \( H(x) = F(x) + x \). \( F(x) \) learns the residual value, which largely reduces the learning complexity.

Each component in the block has a 1D-convolutional layer, a batch normalization layer, and a leaky ReLu layer. Batch Normalization layer normalizes the output of the convolutional layer to make the signal of tensors more outstanding.

Taking an element list as input, the model can analyze the relationship and generate some global features after training. As Figure 2 shows, the input tensor, and the output tensor are 1-D tensors with an element dimension. In the input tensor, each line is an element. The input shows the visual information and the data information, while the output presents which group it belongs to. The information of the visual and data information of each element is encoded in the input tensor. Table 1 presents how to encode the attributes into tensors. Basically, the quantitative attributes are normalized to a range of [0, 1], the categorical attributes are encoded using the one-hot method with a MAX channel number. The output of the model is a list of several features, and each feature classifies the elements into several types.
The value of State A is higher than State B in 2017.

There is a vast space for designing charts. Accordingly, the machine learning model needs a large amount of data to cover the possible examples, replacing the objects and relations in the description, we craft some templates for description generation, following the main idea from Wilbik et al. [17]. The type of feature decides the sprite of the description. Then the description is synthesized by our method. Bottom Right: Underlying data in the tabular form is corresponding to the chart on the left. When users upload a chart in SVG format, after the visual information parsed, the encoding are extracted, and the data information is recovered, Relevant caption would be generated automatically.
5 Evaluation

After the training procedure, we deploy the pre-trained model into our demo system as shown in Figure 3. Figure 4 shows the captions generated from some charts, which verify the effectiveness of the model in a sense. As there is no ready-to-use public tool for chart summarization or baseline dataset to the best of our knowledge, we evaluate our approach through carefully-designed user interviews based on the interactive system.

5.1 Survey

We recruit 15 volunteers for interviews, including a data scientist, an info-graphics designer, a UI designer, an editor, an automation technician, two IT engineers, five college students and three professors from different disciplines like computer science, Chinese language, etc.

In the second part of the survey, participants are required to make simple captions freely as well as to write descriptive sentences under certain constraints like emphasizing trends, comparisons, extremes, and so on. The simple caption represents the fact regarded as most important in the chart, while the other requirement helps us to find out the similarities and differences between the machine-generated captions and manual captions. The summarized result is shown in Table 3. From a statistical perspective, the result of our approach highly accords with most people in the study. However, compared with these manually made captions, some kinds of data facts are neglected by our model. For instance, for the line chart used in the study, three of the fifteen participants describe the wild fluctuation spontaneously while the machine missed. Moreover, in the case of bar charts, our system scores the descriptive sentence about extremes low, yet two people choose the extreme as the prime fact. This is relevant to our crowd-sourcing data collecting strategies, which results in the machine’s preference for the majority.

5.2 Questionnaire

To quantitatively evaluate the captions produced by our approach, we send out online questionnaires and receive 50 valid responses. A majority of the responses come from college students in different disciplines, but there are also software engineers, chemistry engineers, product managers, college professors, librarians, etc.

5.2.1 Design

We mainly assess the effectiveness of our approach from the following five progressive aspects. Moreover, statements in the questionnaire are noted in italic. Participants are required to give a five-order-magnitude rating of their opinions for each statement, where 5 stands for “strongly agree”, 4 for “agree”, 3 for “neutral”, 2 for “disagree”, and 1 for “strongly disagree”.

1. Comprehensibility: The degree to which the user understands the caption. I understand what the caption means.
2. Correctness: The correctness of the content described by the caption. All the generated facts in the caption are correct.
3. Significance: The degree to which the caption deserves to be described. All the generated facts in the caption are worth mentioning.
4. Coverage: The degree to which the important facts are described. The caption mentions all the facts that I think is important in the chart.
5. Helpfulness: To what extent the caption can help the users. The caption provide some information that helps me better understand the chart.

Based on the criteria mentioned above, our current implemented learning model focuses on the Ordinal-Category-Quantity (OCQ) as well as the Quantity-Quantity(QQ) data. There are ten charts in all, including group bar charts, stacked bar charts, line charts, and scatter plots. At the end of the questionnaire, we leave a comment box for any reviews.

5.2.2 Feedback

After collecting the questionnaires and doing analysis, we gain more insights into both the advantages and drawbacks of our method. The summary is shown in Figure 5. From the bar charts of Figure 5 with generally positive attitudes, we may affirm our approach.

Comprehensibility: Everyone finds captions easy to understand and give ratings of five points to approximately 90 percent of the charts. Therefore, the average rating of this criterion reaches surprising 4.8. A freshman of liberal arts indicated that it was acceptable to see such expressions from a machine, and she thought it was good enough.

Correctness: This criterion also gets full recognition and scores high. In most cases, participants find the description consisting of
the chart feature. An engineer from an IT enterprise said that the descriptions about trends were precise, and the judgments of outliers were consistent with his opinion. However, he supposed the caption would be more accurate if there were more details.

**Significance:** About 60 percent of captions are given strong applause for their significance. A junior of data science stated that he found most captions discern the crucial features.

**Coverage:** Participants largely agree that the overall sentences well capture the major features of the charts. A graduate of computer science said that it seemed that the computer could read his mind, and what he thought of were all included in the captions.

**Helpfulness:** Attitudes towards this criterion of the participants vary. On the one hand, some agree that captions help them read the chart with guidance. While on the other hand, some doubt that the captions help little since they are easily readable facts which highly accord with human perceptions.

**Others:** Most participants are satisfied with the captions. In particular, one participant shows great interest in the inner mechanism and want to see the effect of even more complicated charts. However, there exist some complaints about the tedious stereotype. A senior of journalism said that some sentences picked up overmuch exact numbers, whereas human never spoke in that way, and that it would be better if the machine could make a summary. Also, a junior student in history major explained that he was tired of the monotonous sentences with words like “increase” and “unusual drops”.

These opinions reveal the problem of lacking corpus in our current implementation, more or less. Notwithstanding, improvements could be made by enlarging the alternative word database and predefined rules for the feature-description mapping. For instance, adjectives like “greatly”, “sharply”, and “largely” could be used to describe the relatively significant change of trend. Moreover, switching sentences with the same sense would improve the diversity of the expression.

Also, we received suggestions to join closely related sentences together because separated sentences may bring bias to the analyst. Another participant also mentions that a well-organized paragraph, i.e., not individual sentences, would make the combination of the chart and caption easier for the audience to read.

### 6 Conclusion & Future Works

In this work, deep neural networks are employed to construct a learning scheme to achieve the task of caption generation for charts, in which the usage of 1-D residual neural networks (Res-Nets) well shows the ability to extract the feature from the charts. Our model performs well in the feature extraction task, which presents the complicated relationship of data elements. The pre-trained model is deployed to a demo system, where the caption generation based on an uploaded vector-based chart is allowed. The presented results and positive response from the user study demonstrate the effectiveness of our proposed approach.

In the future, we would like to move forward to improve the language quality, the descriptive ability, and the chart type coverage of the captions generated through our approach. This includes a more intelligent caption pipeline, a more natural way to generate language, and broader coverage of the dataset together. We hope to seek research opportunities in the recent progress of natural language processing. Additionally, we regard it as an important direction to explore how the caption and the corresponding chart can be combined, supporting a better experience in exploratory data analysis.

### Acknowledgments

The authors thank the anonymous reviewers for their valuable comments. This work is supported by NSFC No. 61872013 and 61672055. This work is also supported by the Okawa Research Grant.

### References


