Stanford CS 329X - Human-Centered NLP

Lecture 6: Guest Lecture: Interactive Visualization

4/19/23

Lecturer: Sherry Wu
Scribe: Rodrigo Nieto

Readings: N / A.

1 Introduction

Explanations in Human-AI Interaction are effectively communicated to human decision-makers through interfaces Mucha et al. [2021]. The primary function of explanations is to facilitate learning better mental models for how events come about and model visualization is a key form of communication. In addition, people have a preference for certain forms of visualizations even if the underlying information is the same Mucha et al. [2021]. Other than different forms of visualization, having interactive or static interfaces is also another point of consideration as interactive interfaces tend to improve comprehension but it might not be worth the time trade-off, and studies have shown that more interactive models do not make people trust the model more. Therefore, the science of visualization and its relation with HCNLP is itself a key study in order to make the information intuitive and accessible for the reader.

2 High-Level Parameter Map

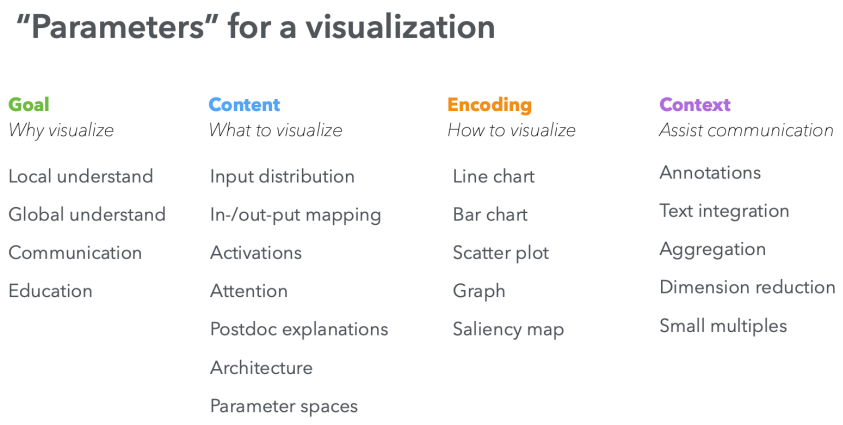

Figure 1: A high-level overview of the "parameters" for a visualization

The parameters for visualization can be partitioned into the following categories: *Goal, Content, Encoding, and Content*. As can be seen in Figure 1, each overall category is context dependent

and offers many different options when considering visualization. Each section will be briefly explored and explained.

3 Goal

3.1 Local understand

The analysis of a local understanding, known as **local interpretability**, involves the understanding of why NLP make their local predictions. For instance, "Which specific token is important when outputting a prediction?"

3.2 Global understand

The analysis of a global understanding, known as **global interpretability**, involves understanding what specific patterns the model has learned in general. For instance, after such an analysis, one might revise the high-level architecture of the model or the dataset. Global understanding also usually involves contrasting outputs with inputs.

3.3 Communication

The process of conveying a certain message (e.g. observations on models) to others. Visualization should be tied to the communication goal. Given a task (e.g. track data and model iterations), the visualization (e.g. bar chart where two different iterations are overlaid) should work in harmony et al. [2020a].

3.4 Education

The process of teaching intuitions and information to a general audience, junior students, etc. These visualizations are usually part of a larger tutorial / example set and are closely integrated with the rest of the text.

4 Content

4.1 Input distribution

A common form of visualizing an input distribution is often through a data map et al. [2020b]. When considering a data map, we often look at the confidence and variability: the means and standard deviation of the gold label probabilities, predicted for each example across training epochs et al. [2020b]. High variability promotes generalize to out-of-distribution (data that is different from the data that the model was trained on) test sets, with little or no effect on in-distribution (distribution of data that the model is trained on) performance. For more context, input data that

has low variability and high confidence play an important role in model optimization. Although this data is not as critical for in-distribution or out-of-distribution performance, without these low variability, high confidence data, training could fail to converge. On the other hand, low variability and low confidence often correspond to labeling errors. Check Figure 2 for more information.

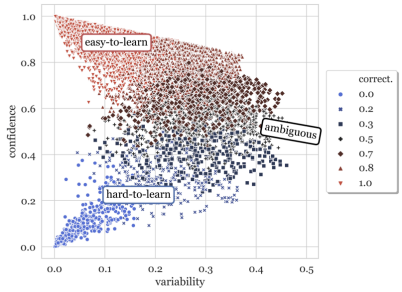

Figure 2: A data map showing the relationship between confidence and variability. et al. [2020b].

4.2 In-/out-put mapping

Input and output mapping is very common for global understanding in order to draw comparisons and contrasts on how different inputs will lead to different outputs. Input and output mapping can also be used in conjunction with analyzing the input distribution to see which kinds of inputs are easier to learn and getting a better sense of which inputs lead to more effective outputs for training, as can be seen in Figure 2.

4.3 Activations

One could inspect neuron firings inside a deep neural network to reveal the complementary and compositional roles that can be played by individual neurons and groups of neurons. Furthermore, factor analysis can be done by decomposing the matrix holding the activation values of Feed-Forward Neural Network neurons using Non-negative Matrix Factorization Alammar [2022]. This technique can be used to analyze the entire network, a single layer, or groups of layers.

4.4 Attention

In Transformers, we can directly model relationships between words in a sentence regardless of their respective positions. For instance, given a certain token in a sequence, we can visualize the self-attention scores for all the tokens in the sequence given our current token Alammar [2022]. Check Figure 3 for an example.

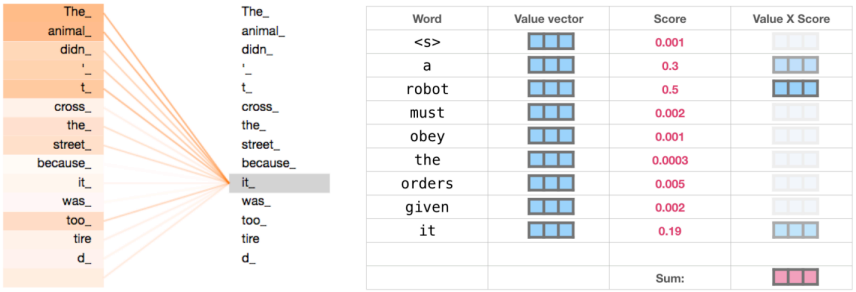

Figure 3: A display of word relationships based on self-attention scores. Alammar [2022].

4.5 Postdoc explanations

Many, such as Mucha et al. [2021], conclude that the exact form of visually representing explanations is relevant for the design of explanations in Human-AI interactions. The study of **visual encoding** can be defined as assigning data fields to visual channels (x, y, color, shape, size, ..) for a chosen graphical mark type (point, bar, line, ...), while also choosing the appropriate encoding parameters (log scale, sorting, ...) and data transformations (bin, group, aggregate, ...). We can partition these data field types into the following categories: **Nominal, Ordered, and Quantitative**. Nominal data field types are often labels and categories with no hierarchical order and are either the same or distinct, such as comparing apples and oranges. Ordered data field types have a clear sense of order and hierarchy, so in addition to different categories being distinct, some categories are better than others (e.g. letter grades on an exam). Quantitative data field types also have hierarchy but in addition can also be interval based (e.g. comparing dates of time of location coordinates) or ratio based (e.g. physical measurements such as length, mass, and temperature). Figure 4 indicates the forms of visual encoding channels that tend to be more effective across data types.

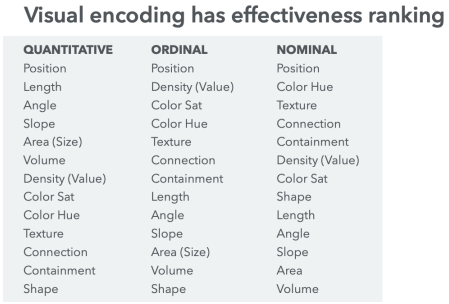

Figure 4: A tentative map of visual encoding channels for certain data types. Some channels are only effective in limited cases Heer.

4.6 Architecture

We may want to also visualize the model architecture to explain the idiosyncrasies or general understanding of the model. For instance, we could have an interactive interface for explaining the Transformer architecture to improve the viewer's comprehension of the Transformer's functionality. For example, check Figure 5.

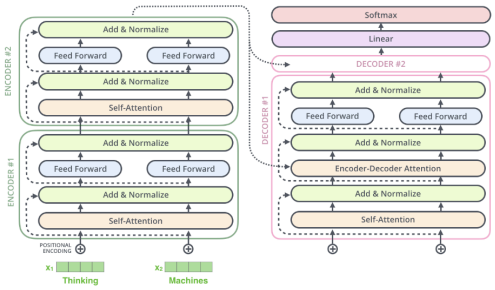

Figure 5: An example of the Transformer architecture described as 2 stacked encoders and decoders Alammar [2022].

4.7 Parameter spaces

With parameter spaces, one can intuitively demonstrate the show of the effects of certain parameters through methods such as dynamic visualization Goh [2017]. They can show how parameters change together because we are rarely interested in a single parameter in isolation. When possible, we want to allow exploration of the possibility space created by multiple parameters Alammar [2022].

5 Encoding

Because people have limited attention spans, they should be given a high-level summary first, before they tailor the visualization based on their interests and knowledge. This also implies that the visualizations themselves should also be sensitive to context and scale. Scalability in particular is a challenge because using more than 5 colors or oscillating colors in a graph can often be overwhelming and may require smoothing.

5.1 Line chart

Line charts can often provide a deeper understanding of the model structure compared to other encoding forms. However, more line charts can often entail more information to interpret, so it is often context-dependent whether it is the most appropriate form.

5.2 Bar chart

Bar charts are a view useful and typical way of displaying data and are particularly common for displaying exact numbers (quantitative).

5.3 Scatter plot

Scatter plots are also a very common form of displaying data that could be useful for temporal trends, data distributions, and especially quantitative data.

5.4 Graph

Graphs can be used in HCNLP in a wide number of contexts and do not have as many data type restrictions as other graphs do. For instance, in order to display a weighted nearest neighbor output, it might be more useful to use a graph to emphasize the shapes and forms of output.

5.5 Saliency map

Saliency maps are a common method used to highlight the most important or visually interesting parts of an image, similar to a heat map. Check Figure 6 as an example.

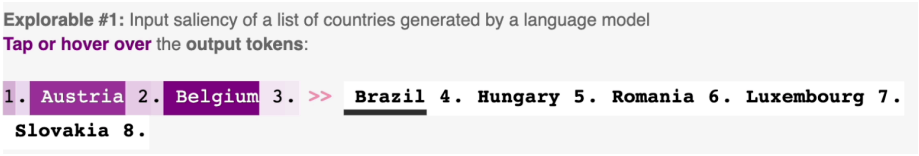

Figure 6: An example of saliency map listing the dependent words for each token Alammar [2022].

6 Context

6.1 Annotations

Annotations help the reader orient themselves by pointing out examples of patterns and important elements.

6.2 Text integration

Text integration is to describe concepts in text, using thoughtful layout and consistent use of color to link their representations visually.

6.3 Aggregation

Aggregation is a common technique to aggregate multiple charts into a singular one with fixed dimensions. For instance, we may be interested in a temporal trend, and use a line chart (vs bar charts or scatter plots), where one dimension is fixed to be the years, where the top words can be annotated to clarify the aggregation.

6.4 Dimension reduction

When we dont have dimensions we can define clearly, we rely on automated methods that project nD data to (not as interpretable) 2D or 3D for viewing. Dimensionality reduction methods are often used to interpret and sanity check high-dimensional representations fit by ML method. Dimensionality are used to aid interpretation, but are also subject to their own interpretation issues. We will anaylyze a method DR methods that each have their own trade-offs:

- **Principal Components Analysis (PCA):** Linear transformation of basis vectors and then ordered by the amount of data variance they explain Heer. First we mean-center the data, we find the orthogonal basis vectors that maximize the data variance, and we plot the data using the top vectors. These linear transformations scale and rotate the original space, where we can preserve the global distances. This distorts the space, thus the trade-off of the preservation of global structure is for emphasizing local neighborhoods.

- **t-Dist. Stochastic Neighbor Embedding (t-SNE)** Probabilistic interpretation of distance, then we optimize the positions Heer. More specifically, we define a model probability **P** of one point and choose another as its neighbor in the original space, using a Gaussian distribution as the distance between points. Thus nearer points have a higher probability than distant ones. Then we define a similar probability **Q** in the lower dimensional embedding space, using a Student's t distribution, which is heavy tailed, allowing distant points to be even further apart. Finally, we optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the **P** and **Q** distributions: $KL(P \mid\mid Q)$. However, it should be noted that we cannot and should not judge relative sizes of clusters in a t-SNE plot.

- **Uniform Manifold Approx. Projection (UMAP)** Identify the local manifolds, then stitch them together Heer. Here we will form weighted nearest neighbor graphs, then layout the graph in a manner that balances the embedding of local and global structures.

6.5 Small multiples

Small multiples can be defined as multiple related charts that share the same scale and axis in order to compare faceted patterns.

References

Jay Alammar. Interfaces for explaining transformer language models. 2022.

Fred Hohman et al. Understanding and visualizing data iteration in machine learning. *CHI Conference on Human Factors in Computing Systems*, 2020a.

Swabha Swayamdipta et al. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *EMNLP*, 2020b.

Gabriel Goh. Why momentum really works. *Distill 2.4*, 2017.

Jeffrey Heer. *UW CS512 Visualization*.

Henrik Mucha, Sebastian Robert, Rudiger Breitschwerdt, and Michael Fellmann. Interfaces for explanations in human-ai interaction: Proposing a design evaluation approach. *CHI Conference on Human Factors in Computing Systems*, 2021.