Guest lecture for CS 329X: Human-Centered NLP
Model Visualization

Sherry Tongshuang Wu
Human Computer Interaction Institute
@tongshuangwu / sherryw@cs.cmu.edu
You should cover:
A quick review on motivation and your project objective
Your method and result
Some discussion on what you learned from your project (limitation, implication for future work, how it could be done differently, etc.)

You will be graded based on:
Presentation clarity, project completeness (an estimation of the effort you put in), and thoughtfulness
Overview

Key things to consider in model visualizations

Common techniques for getting the information to visualize

Local feature attribution

More global dimensionality reduction

And their visual encodings:

Why certain visualization is more effective than others

Key visual encoding channels for different kinds of information
“We found there are significant effects between treatments. We conclude that the exact form of visually representing (LIME) explanations is relevant for the design of explanations in Human-AI interactions.”

Why do people have a preference when it’s very much the same underlying information?

"Although the interactive approach is more effective at improving comprehension, it comes with a trade-off of taking more time."

Why does the interactive approach improve comprehension?

People have studied VIS x AI extensively!

Example interactive visualization articles that explain general concepts and communicate experimental insights when playing with AI models. (a) A Visual Exploration of Gaussian Processes by Göttler, Kehlbeck, and Deussen (VISxAI 2018); (b) What Have Language Models Learned? by Adam Pearce (VISxAI 2021); (c) What if we Reduce the Memory of an Artificial Doom Player? by Jaunet, Vuillemot, and Wolf (VISxAI 2019); (d) Comparing DNNs with UMAP Tour by Li and Scheidegger (VISxAI 2020); (e) The Myth of the Impartial Machine by Feng and Wu (Parametric Press); (f) FormaFiuens Data Experiment by Strobelt, Phibbs, and Martin.

IEEE VIS Workshop: Visualization for AI Explainability
Why do we want to do visualization?

We use visualization to make the information more intuitive and accessible.

**Local interpretability**: Understand why NLP models are making their (local) predictions – which specific token is important?

**Global interpretability**: Get insights into what the model have learned in general.

*Debugging*: Identify potential problems and errors in the model.

**Communication**: Convey certain message (e.g., observations on models) to others.

**Education**: Teach intuitions and information to general audience, junior students, etc.
"Parameters" for a visualization

<table>
<thead>
<tr>
<th>Goal</th>
<th>Content</th>
<th>Encoding</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td>Education</td>
<td>Attention</td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Postdoc explanations</td>
<td>Saliency map</td>
<td>Small multiples</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter spaces</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Parameters” for a visualization

<table>
<thead>
<tr>
<th><strong>Goal</strong></th>
<th><strong>Content</strong></th>
<th><strong>Encoding</strong></th>
<th><strong>Context</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td>Education</td>
<td>Attention</td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Postdoc explanations</td>
<td>Saliency map</td>
<td>Small multiples</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter spaces</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Encoding: Saliency Map

To highlight the most important or visually interesting parts of an image. Saliency maps are commonly used in CV and NLP to identify regions of interest within a document, image or video.

**Explorable #1:** Input saliency of a list of countries generated by a language model

Tap or hover over the output tokens:

Input saliency

Similar information is available across various tasks.

Figure 4: Explaining an image classification prediction made by Google’s Inception neural network. The top 3 classes predicted are “Electric Guitar” \((p = 0.32)\), “Acoustic guitar” \((p = 0.24)\) and “Labrador” \((p = 0.21)\)

**Vanilla gradient**: Approximate the importance of each token, using the gradient of the loss with respect to each token (computed by back-propagating to the input layer).

“For every amount you change this token, I change the output probability of the class/token this much”

**Self-attention**: In Transformers, we can directly model relationships between words in a sentence, regardless of their respective position.

<table>
<thead>
<tr>
<th>Word</th>
<th>Value vector</th>
<th>Score</th>
<th>Value X Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;s&gt;</td>
<td>[ ]</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>[ ]</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>robot</td>
<td>[ ]</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>must</td>
<td>[ ]</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>obey</td>
<td>[ ]</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>[ ]</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>orders</td>
<td>[ ]</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>given</td>
<td>[ ]</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>it</td>
<td>[ ]</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Sum:</td>
<td>[ ]</td>
<td></td>
<td>[ ]</td>
</tr>
</tbody>
</table>
Compute feature attribution using...

**LIME**: Compute local linear approximation of the model's behaviour

"While the model may be very complex globally, it is easier to approximate it around the vicinity of a particular instance."

Look at model's predictions for a bunch of nearby inputs. Closer points are more important than further points. Fit a linear model. Its weights are the feature importances.

The movie is mediocre, maybe even bad.

The movie is mediocre, maybe even bad. **Negative** 98.0%
The movie is mediocre, maybe even bad. **Negative** 98.7%
The movie is mediocre, maybe even bad. **Positive** 63.4%
The movie is mediocre, maybe even bad. **Positive** 74.5%
The movie is mediocre, maybe even bad. **Negative** 97.9%

Reflection: Same visualization different computation

Essentially we end up with a score on each token, seems intuitive to use consistent visualization if we are comparing their algorithms.

Reflection: **Different visualization same computation**

If we fix the algorithm and change the visualization, does that come with any effect? Do you have a preference and why?

Prior work shows that people do have preferences.

“We found there are significant effects between treatments. We conclude that the exact form of visually representing (LIME) explanations is relevant for the design of explanations in Human-AI interactions.”

Visual encoding has effectiveness ranking

**Visual encoding**: Assign **data fields** to **visual channels** \((x, y, \text{color}, \text{shape, size,} \ldots\)\) for a chosen graphical mark type (point, bar, line, \ldots). Also choose appropriate encoding parameters (log scale, sorting, \ldots) and data transformations (bin, group, aggregate, \ldots)

**Data field types:**

- **Nominal (labels or categories)**: Fruits: apples, oranges, \ldots
  
  Operations: \(=, \neq\)

- **Ordered**: Quality of meat: Grade A, AA, AAA Q
  
  Operations: \(=, \neq, >, <\)

- **Quantitative - Interval**: Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
  
  Operations: \(=, \neq, >, <, -\)

- **Quantitative - Ratio (zero fixed)**: Physical measurement: Length, Mass, Temp, \ldots
  
  Operations: \(=, \neq, >, <, -, \%\)
Visual encoding has effectiveness ranking

<table>
<thead>
<tr>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Sat</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color Hue</td>
<td>Connection</td>
</tr>
<tr>
<td>Area (Size)</td>
<td>Texture</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Connection</td>
<td>Density (Value)</td>
</tr>
<tr>
<td>Density (Value)</td>
<td>Containment</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Sat</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
Visual encoding has effectiveness ranking

<table>
<thead>
<tr>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position</strong></td>
<td><strong>Position</strong></td>
<td><strong>Position</strong></td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Saturation</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Hue</td>
<td>Color Saturation</td>
</tr>
<tr>
<td>Slope</td>
<td>Texture</td>
<td>Length</td>
</tr>
<tr>
<td>Area (Size)</td>
<td>Connection</td>
<td>Angle</td>
</tr>
<tr>
<td>Volume</td>
<td>Containment</td>
<td>Slope</td>
</tr>
<tr>
<td>Density (Value)</td>
<td></td>
<td>Area (Size)</td>
</tr>
<tr>
<td><strong>Color Saturation</strong></td>
<td>Color Hue</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Texture</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Connection</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Containment</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td></td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td></td>
<td>Volume</td>
</tr>
</tbody>
</table>

Some visual encoding channels tend to be more effective across data types; some channels are only effective in limited cases.

Slides: Jeffrey Heer, UW CSE 512 Visualization
Reflection: Different visualization same computation

If you prefer exact numbers (**quantitative**), bar charts make sense (**length encoding**). But most of time we only care about the rough relative importance (for glance – **ordinal**!), which makes color more effective.

Cautious! Saliency map leads to cognitive bias.
Cautious! Saliency map leads to cognitive bias.

### “Parameters” for a visualization

<table>
<thead>
<tr>
<th>Goal</th>
<th>Content</th>
<th>Encoding</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td>Education</td>
<td><strong>Attention</strong></td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Postdoc explanations</td>
<td><strong>Saliency map</strong></td>
<td>Small multiples</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter spaces</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Parameters” for a visualization

<table>
<thead>
<tr>
<th>Goal</th>
<th>Content</th>
<th>Encoding</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td>Education</td>
<td>Attention</td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Postdoc explanations</td>
<td>Saliency map</td>
<td>Small multiples</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter spaces</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Neuron Activations & Factor Analysis

Inspect neuron firings inside deep neural networks can reveal the complementary and compositional roles that can be played by individual neurons, and groups of neurons. **Compared to saliency maps:** Deeper understanding of the model structure. But, more information to interpret (usually end up “forcing” meanings onto factors).

Neuron Activations & Factor Analysis

Factor analysis is done by decomposing the matrix holding the activations values of Feed-Forward Neural Network (FFNN) neurons using Non-negative Matrix Factorization. It can be used to analyze the entire network, a single layer, or groups of layers.

Neuron Activations & Factor Analysis

Useful demo case: DistilGPT2 reacts to XML. Shows a clear distinction of factors attending to different components of the syntax.

New-lines
Labels of tags, with higher activation on closing tags
Indentation spaces
The '<' (less-than) character starting XML tags
The large factor focusing on the **first token**. Common to GPT2 models.
Two factors tracking the '>' (greater than) character at the end of XML tags
The text inside XML tags
The '</' symbols indicating closing XML tag
Encoding / viz.: Important techniques

“Overview first, details on-demand”: People have limited attention span. They should be given a high level summary first, before they tailor the viz based on their interest and knowledge.

Overview: line charts + the max color for each token; Detail: coloring tokens using specific factors.

Small multiples: Multiple related charts that share same scale and axis, to compare faceted patterns.

Linked views: Set of coordinated visualizations that are connected such that interactions in one visualization affect the others. Help users explore and analyze data from multiple perspectives.

Text integration: When describing concepts in text, link their representations visually via e.g., thoughtful layout and consistent use of color.

---

Explorable: Ten Activation Factors of XML

Tap or hover over the sparklines on the left to isolate a certain factor

Factorizing neuron activations in response to XML (that was generated by an RNN from XML) into ten factors results in factors corresponding to:

1. New lines
2. Labels of tags, with higher activation on closing tags
3. Indentation space
4. The '<' character starting XML tags
5. The large factor focusing on the '<' character at the end of XML tags
6. Two factors tracking the '>' (greater than) character at the end of XML tags
7. The space inside XML tags
8. The '<' symbols indicating closing XML tag
Is the same overview always useful?

**Scalability** challenge!

> 5 color is usually overwhelming, oscillating colors (or lines) are also overwhelming. Sometimes **smoothing** is helpful.
Font property - Size
Title: A meta analysis of birth origin effects on reproduction in diverse captive environments
Abstract: Successfully establishing captive breeding programs is priority across diverse industries to address food security demand for ethical laboratory research animals and prevent extinction differences in reproductive success due to birth origin may threaten the long term sustainability of captive breeding programs. Our meta analysis examining effect sizes from species of invertebrates, fish, birds and mammals shows that overall captive born animals have decreased odds of reproductive success in captivity compared to their wild born counterparts. The largest effects are seen in commercial aquaculture related to conservation or laboratory settings and offspring survival and offspring quality were the most sensitive traits although somewhat weaker trend reproductive success in conservation and laboratory research breeding programs is also in negative direction for captive born animals. Our study provides the foundation for future investigation of non genetic and genetic drivers of change.

Font property - Luminance
Title: A meta analysis of birth origin effects on reproduction in diverse captive environments
Abstract: Successfully establishing captive breeding programs is priority across diverse industries to address food security demand for ethical laboratory research animals and prevent extinction differences in reproductive success due to birth origin may threaten the long term sustainability of captive breeding programs. Our meta analysis examining effect sizes from species of invertebrates, fish, birds and mammals shows that overall captive born animals have decreased odds of reproductive success in captivity compared to their wild born counterparts. The largest effects are seen in commercial aquaculture related to conservation or laboratory settings and offspring survival and offspring quality were the most sensitive traits although somewhat weaker trend reproductive success in conservation and laboratory research breeding programs is also in negative direction for captive born animals. Our study provides the foundation for future investigation of non genetic and genetic drivers of change.

Additional Mark - Circle Area
Title: A meta analysis of birth origin effects on reproduction in diverse captive environments
Abstract: Successfully establishing captive breeding programs is priority across diverse industries to address food security demand for ethical laboratory research animals and prevent extinction differences in reproductive success due to birth origin may threaten the long term sustainability of captive breeding programs. Our meta analysis examining effect sizes from species of invertebrates, fish, birds and mammals shows that overall captive born animals have decreased odds of reproductive success in captivity compared to their wild born counterparts. The largest effects are seen in commercial aquaculture related to conservation or laboratory settings and offspring survival and offspring quality were the most sensitive traits although somewhat weaker trend reproductive success in conservation and laboratory research breeding programs is also in negative direction for captive born animals. Our study provides the foundation for future investigation of non genetic and genetic drivers of change.

Additional Mark - Background Color Intensity
Title: A meta analysis of birth origin effects on reproduction in diverse captive environments
Abstract: Successfully establishing captive breeding programs is priority across diverse industries to address food security demand for ethical laboratory research animals and prevent extinction differences in reproductive success due to birth origin may threaten the long term sustainability of captive breeding programs. Our meta analysis examining effect sizes from species of invertebrates, fish, birds and mammals shows that overall captive born animals have decreased odds of reproductive success in captivity compared to their wild born counterparts. The largest effects are seen in commercial aquaculture related to conservation or laboratory settings and offspring survival and offspring quality were the most sensitive traits although somewhat weaker trend reproductive success in conservation and laboratory research breeding programs is also in negative direction for captive born animals. Our study provides the foundation for future investigation of non genetic and genetic drivers of change.

“Parameters” for a visualization

Goal
Why visualize

Local understand
Input distribution
In-/out-put mapping
Activations
Attention
Postdoc explanations
Architecture
Parameter spaces

Content
What to visualize

Encoding
How to visualize
Line chart
Bar chart
Scatter plot
Graph
Saliency map

Context
Assist communication
Annotations
Text integration
Aggregation
Dimension reduction
Small multiples
**“Parameters” for a visualization**

<table>
<thead>
<tr>
<th>Goal</th>
<th>Content</th>
<th>Encoding</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td>Education</td>
<td>Attention</td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Postdoc explanations</td>
<td>Saliency map</td>
<td>Small multiples</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter spaces</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Parameters” for a visualization

Goal
Why visualize

Local understand

Global understand
Communication
Education

Content
What to visualize

Input distribution
In-/out-put mapping
Activations
Attention
Postdoc explanations
Architecture
Parameter spaces

Encoding
How to visualize

Line chart
Bar chart
Scatter plot
Graph
Saliency map

Context
Assist communication

Annotations
Text integration
Aggregation
Dimension reduction
Small multiples
Key: Project info onto readable dimensions

**Global understanding** usually involves contrasting outputs with inputs.

When we have many outputs, good to **map them out on dimensions we care about**.

**Color encoding** helps highlight the contrast.

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

Some amount of **aggregation** is also important. Here we are interested in the temporal trend, which is more suitable for line chart (vs. bar chart or scatter plots). As a result, one dimension is fixed to be year, and top words are annotated on the side.
Instances that a model always predicts correctly are different from those it almost never does, or those on which it vacillates.

How to get data map? **Confidence** and **Variability**: the mean and standard deviation of the gold label probabilities, predicted for each example across training epochs.
low variability, high confidence; play an important role in model optimization. Not as critical for ID or OOD performance, but without any such instances, training could fail to converge.

Low variability, low confidence; often correspond to labeling errors.

Color encoding: Continuous, 2-D color hue for soft categorization.

High variability; Promotes generalization to out-of-distribution test sets, with little or no effect on in-distribution (ID) performance.
Projection through Dimensionality Reduction

When we don’t have dimensions we can define clearly, we rely on automated methods that project nD data to (not as interpretable) 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by ML methods.

DR methods are used to aid interpretation, but are also subject to their own interpretation issues!

Different DR methods make different trade-offs: for example to preserve global structure (e.g., PCA) or emphasize local structure (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

Slides: Jeffrey Heer, UW CSE 512 Visualization
Dimensionality Reduction Methods

Principal Components Analysis (PCA)
Linear transformation of basis vectors, ordered by amount of data variance they explain.

$t$-Dist. Stochastic Neighbor Embedding (t-SNE)
Probabilistically model distance, optimize positions.

Uniform Manifold Approx. & Projection (UMAP)
Identify local manifolds, then stitch them together.
Projection (1/3): Principal Components Analysis

1. Mean-center the data.
2. Find $\perp$ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

Linear transform: scale and rotate original space.
Lines (vectors) project to lines.
Preserves global distances.
Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

**t-SNE** - probabilistic interpretation of distance

**UMAP** - tries to balance local/global trade-off
Projection (2/3): t-SNE [Maaten & Hinton 2008]

Model probability $P$ of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

Define a similar probability $Q$ in the low-dimensional (2D or 3D) embedding space, using a Student’s $t$ distribution (hence the “$t$-” in “t-SNE”!). The $t$-distribution is heavy-tailed, allowing distant points to be even further apart.

Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the $P$ and $Q$ distributions: $KL(P \parallel Q)$

Projection (2/3): t-SNE [Maaten & Hinton 2008]

Figure 2: A t-SNE projection of the embedding of 74 semantically identical sentences translated across all 6 possible directions, yielding a total of 9,978 steps (dots in the image), from the model trained on English↔Japanese and English↔Korean examples. (a) A bird’s-eye view of the embedding, coloring by the index of the semantic sentence. Well-defined clusters each having a single color are apparent. (b) A zoomed in view of one of the clusters with the same coloring. All of the sentences within this cluster are translations of “The stratosphere extends from about 10km to about 50km in altitude.” (c) The same cluster colored by source language. All three source languages can be seen within this cluster.

Visualization could be misleading

Cluster sizes in a t-SNE plot mean nothing

Non-linear projection (or really, any computation + visualization method) needs to be used with caution.

e.g., t-SNE adapts its notion of “distance” to regional density variations in the data set: it expands dense clusters, and contracts sparse ones, evening out cluster sizes. i.e., **Density equalization happens by design and is a predictable feature of t-SNE.**

As a result we cannot & should not judge relative sizes of clusters in a t-SNE plot.

Distances between clusters might not mean anything

!For more t-SNE pitfalls: https://distill.pub/2016/misread-tsne/
Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

“Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance.” - McInnes et al. 2018
A visual comparison between algorithms

※ ●

Again, small multiples and linked views :)
“Parameters” for a visualization

**Goal**
Why visualize

**Global understand**
Local understand
Communication
Education

**Content**
What to visualize

**Input distribution**
In-/out-put mapping
Activations
Attention
Postdoc explanations
Architecture
Parameter spaces

**Encoding**
How to visualize

Line chart
Bar chart
Scatter plot
Graph
Saliency map

**Context**
Assist communication
Annotations
Text integration
Aggregation
Dimension reduction
Small multiples

**Goal**
Why visualize

**Content**
What to visualize

**Input distribution**
In-/out-put mapping
Activations
Attention
Postdoc explanations
Architecture
Parameter spaces

**Encoding**
How to visualize

Line chart
Bar chart
Scatter plot
Graph
Saliency map

**Context**
Assist communication
Annotations
Text integration
Aggregation
Dimension reduction
Small multiples
Multi-view, interactive interfaces for understanding

We use multi-view interactions because...
Humans have a lot of inter-twined goals that cannot be embedded into a single view.
We have too much information to present at once.
Allow humans to inquiry targeted information.
Suggest information to mitigate human biases.

Best practices for these system designs
(Again) “Overview first, details on-demand”
Integrate into users’ natural developing environments
Explore intuitive interactions
Basic idea: A framework for testing models on nuanced capabilities.

An example of lightweight visual interface

“Overview first, details on-demand”

Different views can be invoked in Jupyter Notebook

Example: CheckList

Ribeiro, Marco Tulio, et al. “Beyond accuracy: Behavioral testing of NLP models with CheckList.” ACL 2020
### Minimum Functionality Test

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Failure Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple negations: negative</td>
<td>42 / 500 = 8.4%</td>
</tr>
<tr>
<td>Simple negations: not negative</td>
<td>66 / 500 = 13.2%</td>
</tr>
<tr>
<td>Simple negations: not neutral is still neutral</td>
<td>492 / 500 = 98.4%</td>
</tr>
<tr>
<td>Simple negations: I thought x was positive, but it was not (should be negative)</td>
<td>11 / 500 = 2.2%</td>
</tr>
<tr>
<td>Simple negations: I thought x was negative, but it was not (should be neutral or positive)</td>
<td>424 / 500 = 84.8%</td>
</tr>
<tr>
<td>Simple negations: but it was not (neutral) should still be neutral</td>
<td>493 / 500 = 98.6%</td>
</tr>
<tr>
<td>Hard: Negation of positive with neutral stuff in the middle (should be negative)</td>
<td>370 / 500 = 74.8%</td>
</tr>
<tr>
<td>Simple Negations</td>
<td>Accuracy</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>not negative</td>
<td>13.2%</td>
</tr>
<tr>
<td>neutral is still neutral</td>
<td>98.4%</td>
</tr>
<tr>
<td>I thought x was positive, but it was not (should be negative)</td>
<td>2.2%</td>
</tr>
<tr>
<td>I thought x was negative, but it was not (should be neutral or positive)</td>
<td>84.8%</td>
</tr>
<tr>
<td>but it was not (neutral) should still be neutral</td>
<td>98.6%</td>
</tr>
</tbody>
</table>

Test Summary

Test: [MFT] on [NEGATION]

Hard: Negation of positive with neutral stuff in the middle (should be negative)

Result: Failure rate on all cases

370/500 = 74.0%

Examples

- I would not say, given that I am from Brazil, that this food was extraordinary.
  - Expect: 0 | Pred: 2 (1.00)

- I would not say, given it’s a Tuesday, that that is a beautiful aircraft.
  - Expect: 0 | Pred: 2 (1.00)

- I would not say, given that I am from Brazil, that the service is wonderful.
  - Expect: 0 | Pred: 2 (1.00)

- I can’t say, given all that I’ve seen.
  - Expect: 0 | Pred: 2 (1.00)
<table>
<thead>
<tr>
<th>Simple Negations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>not negative</td>
<td>13.2%</td>
</tr>
<tr>
<td>not neutral is still neutral</td>
<td>98.4%</td>
</tr>
<tr>
<td>I thought x was positive, but it was not (should be negative)</td>
<td>2.2%</td>
</tr>
<tr>
<td>I thought x was negative, but it was not (should be neutral or positive)</td>
<td>84.8%</td>
</tr>
<tr>
<td>but it was not (neutral) should still be neutral</td>
<td>98.6%</td>
</tr>
<tr>
<td>Hard: Negation of positive with neutral stuff in the middle (should be negative)</td>
<td>74.0%</td>
</tr>
</tbody>
</table>

Test Summary

Test [MFT] on [NEGATION]

Result  Failure rate on all cases

370/500 = 74.0%

Examples

<table>
<thead>
<tr>
<th>Input</th>
<th>Expect</th>
<th>Pred</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would n't say, given that I am from Brazil, that this food was extraordinary.</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>I would n't say, given it's a Tuesday, that is a beautiful aircraft.</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>I would n't say, given that I am from Brazil, that the service is wonderful.</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>I can't say, given all that I've seen.</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Simple Negations: Not Negative</td>
<td>65 / 500 = 13.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Negations: Not Neutral Is Still Neutral</td>
<td>492 / 500 = 98.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Negations: I Thought x Was Positive, But It Was Not (Should Be Negative)</td>
<td>11 / 500 = 2.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Negations: I Thought x Was Negative, But It Was Not (Should Be Neutral Or Positive)</td>
<td>424 / 500 = 84.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Negations: But It Was Not (Neutral) Should Still Be Neutral</td>
<td>493 / 500 = 98.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard: Negation of Positive With Neutral Stuff in the Middle (Should Be Negative)</td>
<td>370 / 500 = 74.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Test Summary**

Test [MFT] on [NEGATION]

Hard: Negation of positive with neutral stuff in the middle (should be negative)

**Result** Failure Rate On All Cases

370/500=74.0%
<table>
<thead>
<tr>
<th>Simple Negations: Not Negative</th>
<th>65 / 500 = 13.2%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Negations: Not Neutral is Still Neutral</td>
<td>492 / 500 = 98.4%</td>
</tr>
<tr>
<td>Simple Negations: I Thought x Was Positive, But It Was Not (Should Be Negative)</td>
<td>11 / 500 = 2.2%</td>
</tr>
<tr>
<td>Simple Negations: I Thought x Was Negative, But It Was Not (Should Be Neutral or Positive)</td>
<td>424 / 500 = 84.8%</td>
</tr>
<tr>
<td>Simple Negations: But It Was Not (Neutral) Should Still Be Neutral</td>
<td>493 / 500 = 98.6%</td>
</tr>
</tbody>
</table>

Hard: Negation of positive with neutral stuff in the middle (should be negative) 370 / 500 = 74.8%
Example: Language Interpretability Tool (LIT)

**Basic idea:** A tool that integrates various global & local explanations.

**An example of multi-view visualization**

Different aspects of information presented in different views

Interactions invoke linked update
“Parameters” for a visualization

<table>
<thead>
<tr>
<th>Goal</th>
<th>Content</th>
<th>Encoding</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td>Education</td>
<td>Postdoc explanations</td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td>Saliency map</td>
<td>Small multiples</td>
</tr>
<tr>
<td></td>
<td>Parameter spaces</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Visualization should be tied to **communication goal**

![Figure 3](image)

**Task:** track data & model iterations.

**Viz:** bar chart, but data from two different iterations overlaid.

---

Hohman, Fred, et al. "Understanding and visualizing data iteration in machine learning." *CHI 2020*
Visualization should be tied to **communication goal**

**Simulating loan decisions for different groups**

Drag the black threshold bars left or right to change the cut-offs for loans. Click on different preset loan strategies.

**Loan Strategy**

- Maximize profit with:
  - **MAX PROFIT**: No constraints
  - **GROUP UNAWARE**: Blue and orange thresholds are the same
  - **DEMOGRAPHIC PARITY**: Same fractions blue / orange loans
  - **EQUAL OPPORTUNITY**: Same fractions blue / orange loans to people who can pay them off

**Blue Population**

<table>
<thead>
<tr>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
</table>

**Goal**: Explore and explain model fairness.

**Viz**: Multiple linked views with color-encoded groups.

**Attacking discrimination with smarter machine learning**
### “Parameters” for a visualization

<table>
<thead>
<tr>
<th><strong>Goal</strong></th>
<th><strong>Content</strong></th>
<th><strong>Encoding</strong></th>
<th><strong>Context</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Why visualize</td>
<td>What to visualize</td>
<td>How to visualize</td>
<td>Assist communication</td>
</tr>
<tr>
<td>Local understand</td>
<td>Input distribution</td>
<td>Line chart</td>
<td>Annotations</td>
</tr>
<tr>
<td>Global understand</td>
<td>In-/out-put mapping</td>
<td>Bar chart</td>
<td>Text integration</td>
</tr>
<tr>
<td>Communication</td>
<td>Activations</td>
<td>Scatter plot</td>
<td>Aggregation</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td>Postdoc explanations</td>
<td>Graph</td>
<td>Dimension reduction</td>
</tr>
<tr>
<td></td>
<td>Architecture</td>
<td>Saliency map</td>
<td>Small multiples</td>
</tr>
</tbody>
</table>
Intuitively demonstrate the show the effects of certain parameters, via dynamic visualization.

We often think of Momentum as a means of dampening oscillations and speeding up the iterations, leading to faster convergence. But it has other interesting behavior. It allows a larger range of step-sizes to be used, and creates its own oscillations. What is going on?

Content: Parameter Space

These visualizations are usually part of a larger tutorial / example set, and are closely integrated with the rest of the text.

Let's see how this plays out in polynomial regression. Given 1D data, \( \xi_i \), our problem is to fit the model

\[
\text{model}(\xi) = w_1 p_1(\xi) + \cdots + w_n p_n(\xi) \quad p_i = \xi \mapsto \xi^{i-1}
\]

to our observations, \( d_i \). This model, though nonlinear in the input \( \xi \), is linear in the weights, and therefore we can write the model as a linear combination of monomials, like:

\[
\begin{align*}
-2.00p_1 & + -2.00p_2 & + 2.00p_3 & + 2.00p_4 & + 2.00p_5 & + -2.00p_6 & = \text{model}
\end{align*}
\]
Show how parameters change together. We rarely are interested in a single parameter in isolation. When possible, allow exploration of the possibility space created by multiple parameters.

More elegantly: Map the space created by 2 parameters and link views to outputs. Annotate regions of parameter space.
Why interaction improves comprehension?

“Although the interactive approach is more effective at improving comprehension, it comes with a trade-off of taking more time.”

Having the control to isolate/combine different variables is important.

Understanding Convolutions

Interfaces for Explaining Transformer Language Models

The Transformer architecture has been powering recent advances in natural language processing. Pre-trained language models (models that use their own output as input to themselves) are widely deployed for tasks like pre-training and then fine-tuning. However, explaining the workings of these models remains a challenge. Recently, researchers have proposed a new method to explain the behavior of Transformer models, called “Interfaces for Explaining Transformer Language Models.”

6th Workshop on Visualization for AI Explainability

The role of visualization in artificial intelligence (AI) gained significant attention in recent years. With the growing complexity of AI models, the critical need for understanding their inner workings has increased. Visualization is potentially a powerful technique to fill such a critical need.

The goal of this workshop is to initiate a call for “explaneable”/“explainable” that explain how AI techniques work using visualization. We believe the VIS community can leverage their expertise in creating visual narratives to bring new insights into the often obfuscated complexity of AI systems.

Probabilities?

Machine learning models express their uncertainty as model scores, but they often lack interpretability. To help with this, we can transform these scores into probabilities for more effective decision-making.

Every dataset communicates a different perspective. When you shift your perspective, your conclusions can shift, too.
Some reflection on the parameters, and VIZ x NLP

The most important thing of visualization is you want to achieve some **goal**, using certain **content**.

There has been 20+ years of study on effective visualization (e.g. line chart better for trend, must be for quantitative values; bar chart better for comparison). Usually once you know your goal, it’s not too hard to find optimal **visualization encodings**.

**Clear legend & textual annotation** is essential.

Importantly, **content** can really be **any information** you can compute and obtain around your model – input, output, all sorts of scores. viz. is a shared topic across data collection/curation, model training and debugging, deployment, and knowledge sharing, and probably shouldn’t be taken for granted :)

---

**Goal**
Why visualize

**Content**
What to visualize

**Encoding**
How to visualize

**Context**
Assist communication
Altair (in my opinion) the best visualization package with various encoding options.

Ecco a viz library for Language Model feature attribution and neuron activations.

Jupyter Widget + React Most typical way to build Notebook-embedded plug-ins.

CohereAI / Jay Alammar has (in my opinion) the most useful visualization for NLP beginners

Distill.pub and PAIR explorable has interactive articles you can play with.

Draco or VizLinter has some quick overview on visualization constraint 101.
Recap

Model visualization can happen at any stage in model
development and deployment.

Visualization encoding changes based on what patterns we are
trying to convey, based on what data.

Most common visualizations overlay information on top of
dimensions we are familiar of (token-wise saliency map); Others
reduce the uninterpretable dimension to some interpretable
number (dimensionality reduction).

More holistic linked views give you more holistic understanding,
but require more effort (to build, and to interact with).