A causal framework for explaining the predictions of black-box sequence-to-sequence models

Scribe: Bill Zhang

March 1, 2019

1 Introduction

Problem: Interpretability of complex models

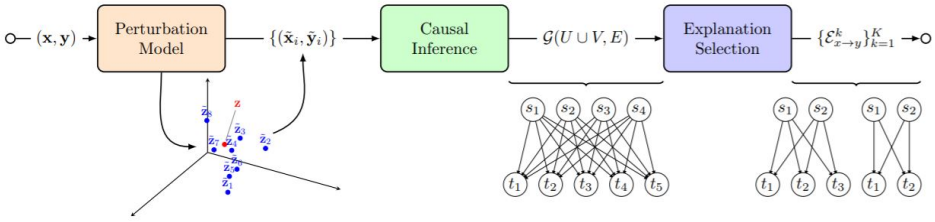
If models are easier to interpret, then we can have more trust in the models, better error analysis, and better model refinement.

Two focuses of problem:

- Model interpretability: explaining the architecture
- Prediction interpretability: explaining particular predictions of model

This paper focuses on improving prediction interpretability with only oracle access to the model generating the prediction.

2 Big Picture Model: SocRat



Black-box model $F : X \to Y$. Every $\mathbf{x} \in X$ and $\mathbf{y} \in Y$ are feature-set representations (so they can be sequences, graphs, images,...). For every input-output pair $(\mathbf{x}, \mathbf{y})$, we want to generate an explanation of $\mathbf{y}$ in terms of $\mathbf{x}$.

Summarize behavior of F around $\mathbf{x}$ as a weighted bipartite graph $G = (V_x \cup$

$V_y, E)$, where weights of edges represent amount of influence and V_x, V_y are elements of $\mathbf{x}$ and $\mathbf{y}$. Finally, take subgraphs $G^k = (V_x^k \cup V_y^k, E^k)$ from G where each subgraph is an explaining component. An explanation is a collection of components G^1 to G^k.

3 Model Details

3.1 Perturbation Model

Goal: obtain perturbed versions of input which are similar with potential changes in order and elements.

Do this with a variational autoencoder (VAE) which perturbs in continuous latent representation rather than directly on inputs. VAE is trained in a way in which semantic content remains similar. Repeatedly sample from VAE and map back to original space to obtain perturbed samples.

3.2 Causal Inference

Goal: Use perturbed input-output pairs in infer causal dependencies between original input and output.

Incorporate all input tokens to predict occurrence of single output token via logistic regression. Also interested in uncertainty of these predictions, which can be done with a Bayesian approache to logistic regression.

Let $\phi_x(\tilde{x}) \in \{0,1\}^{|x|}$ be a binary vector encoding the presence of original tokens $x_1, ..., x_n$ from $\mathbf{x}$ in the perturbed $\tilde{\mathbf{x}}$. Then, estimate a model:

$$P(y_j \in \tilde{\mathbf{y}} | \tilde{\mathbf{x}}) = \sigma(\boldsymbol{\theta}_j^T \phi_x(\tilde{\mathbf{x}}))$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$. Upon completion of this step, we have dependency coefficients between all original input and output tokens $\{\theta_{i,j}\}$.

3.3 Explanation Section

Goal: Make bipartite graph more interpretable by extracting most relevant components of graph.

Similar to graph partitioning problem (NP-complete), except incorporate uncertainty of dependency estimates. Use robust optimization formulation which minimizes worst case cut values, casting as a mixed integer programming problem (MIP).

On a high level, this step returns a set of partitions of a given graph $G = (U, V, E)$. If the dimension of $\mathbf{x}$ and $\mathbf{y}$ are small, then all partitions can be returned. However, for larger pairs, only the top κ partitions are of interest. Each

partition can be ranked based on how self-contained they are; if a partition has few high-valued edges connecting them to other parts of the graph, then they are more self-contained.

For a full detailed explanation of this step, see the paper (the formulas are omitted due to their complexity).

4 Experiments and Results

4.1 Recovering Simple Mappings

Use CMU dictionary of word pronunciations, consisting of 130K word-to-phoneme pairs. Mapping is simple, but not one-to-one or deterministic.

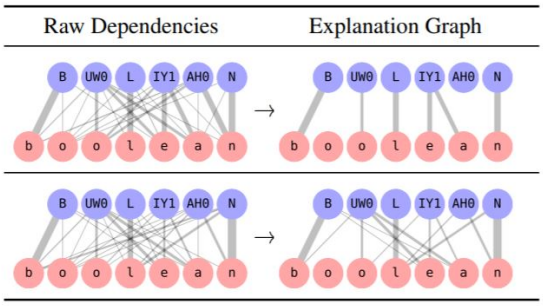


Table 1: Inferred dependency graphs before (left) and after (right) explanation selection for the prediction: *boolean* $\mapsto$ `B UW0 L IY1 AH0 N`, in independent runs with large (top) and small (bottom) clustering parameter k.

4.2 Machine Translation

Three methods of translating English to German:

- Azure's Machine Translation System
- Neural MT model
- Human

Translate same sentences, and explain predictions using SocRat.

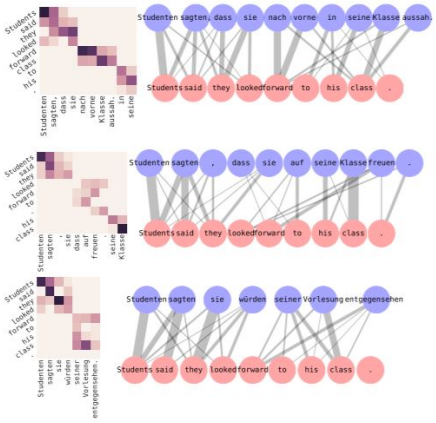


Figure 3: Explanations for the predictions of three Black-Box translators: Azure (top), NMT (middle) and human (bottom). Note that the rows and columns of the heatmaps are permuted to show explanation *chunks* (clusters).

Emphasize that attention scores are not necessary to use SocRat. SocRat may be preferable in some cases because it is more interpretable and can be retrieved in order of relevance.

4.3 Mediocre Dialogue System

Train simple dialogue system on OpenSubtitle corpus, but constain quality (only 2 layers, hidden size 1000, no hyperparameter tuning).

Input	Prediction
What do you mean it doesn't matter?	*I don't know*
Perhaps have we met before?	*I don't think so*
Can I get you two a cocktail?	*No, thanks.*

Table 2: "Good" dialogue system predictions.

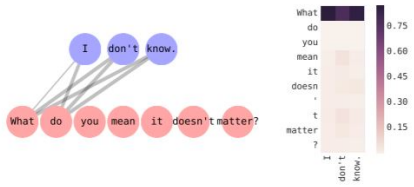


Figure 5: Explanation with $S = 50$ (left) and attention (right) for the first prediction in Table 2.

"Good" predictions can be shown to be memorization using SocRat.

4.4 Bias Detection

Many datasets have biases. Simulate biased dataset with WMT14 English-French dataset. Identify French sentences written in informal register and prepend English translation with "However" to try to induce false association between informal register and the word "However."

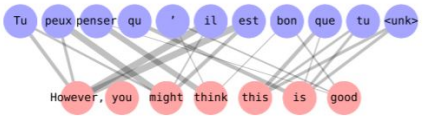


Figure 6: Explanation with $S = 50$ for the prediction of the biased translator.

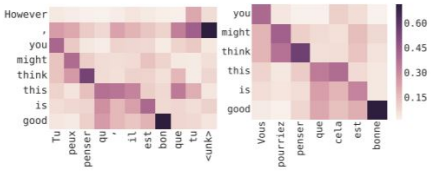


Figure 7: Attention scores on similar sentences by the biased translator.

SocRat detects the association of "However" to "tu" or "peux", which are in informal register sentences, showing that it can detect known biases.

Can also show gender biases; translate gender-neutral sentences into sentences requiring a gender-declined words.

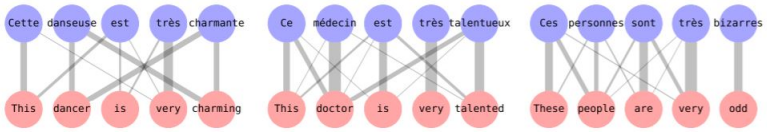


Figure 8: Explanations for biased translations of similar gender-neutral English sentences into French generated with Azure's MT service. The first two require gender declination in the target (French) language, while the third one, in plural, does not. The dependencies in the first two shed light on the cause of the biased selection of gender in the output sentence.

5 Conclusion

- Created model-agnostic framework for prediction interpretability which produces reasonable, coherent, and insightful explanations
- Framework allows for partial-view of black-box systems and shows potential to improve existing systems
- Future work can be done to test effectiveness on other types of data (not just seq2seq tasks)