Evaluating Saliency Methods for Neural Language Models

Shuoyang Ding  Philipp Koehn

NAACL 2021
Online
Saliency Methods

• A specific category of interpretation methods for neural network models.

• Regarding a prediction $y$ made by a model $M$, give a distribution of feature contributions over the input feature set $F$. 
Saliency Methods

The wife checked the son for injuries because

Left-to-right Language Model
Saliency Methods

The wife checked the son for injuries because she...
Saliency Methods

• A specific category of **interpretation methods** for neural network models.

• Regarding a prediction \( y \) made by a model \( M \), give a distribution of feature contributions over the input feature set \( F \).

• Example: \( \frac{\partial y}{\partial x} \) for an input feature \( x \) in set \( F \)
The reality...

<table>
<thead>
<tr>
<th>V</th>
<th>U.S. companies</th>
<th>wanting to expand in Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>U.S. companies</td>
<td>wanting to expand in Europe</td>
</tr>
<tr>
<td>IG</td>
<td>U.S. companies</td>
<td>wanting to expand in Europe</td>
</tr>
</tbody>
</table>

- Different methods can yield different interpretations for the same model prediction
- Which one to trust?
This Work

• Build a benchmark to evaluate how *trustworthy* saliency methods are
Define *Trustworthy*.

- Plausible
- Faithful

Largely from Jacovi & Goldberg (2020)
Define *Trustworthy*. 

- Plausible
- Faithful
Plausibility

• Saliency methods are interpretation methods, so humans should be able to understand them.

• Hence, they should align with basic human intuitions about how model decisions are made.
Plausibility

Yao et al. 2018

*Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions*
Challenge

Lack of ground truth interpretation
However, most people, having been subjected to news footage of the devastated South Bronx, (look/looks...)

- Solution: leverage lexical agreements to build interpretation ground truth
Solution

However, most people, having been subjected to news footage of the devastated South Bronx, (look/ looks...)

• **Cue**: words in the input prefix that could lead the model to predict the correct next word
Solution

However, most people, having been subjected to news footage of the devastated South Bronx, (look/ looks...)

- **Attractor**: words in the prefix that could lead the model to predict the incorrect next word
However, most **people**, having been subjected to news **footage** of the devastated South **Bronx**, (look/looks...)

- **Expected Case**: when the model predicts the correct next word **look**, the interpretation passes the test if:

  \[
  \max \psi(\text{cue}) > \max \psi(\text{attractor})
  \]

  \(\psi\): feature importance/saliency
Solution

However, most people, having been subjected to news footage of the devastated South Bronx, (look/looks...)

- **Alternative Case**: When the model predicts the incorrect next word looks, the interpretation passes the test if:

\[
\max \psi(cue) < \max \psi(attractor)
\]

\(\psi\): feature importance/saliency
Problem

• Lack of **constraints on agreements**
  • prediction of a word could incur multiple agreements at the same time
  • consider the prefix: 
    *At the polling station men and women...*
    • vote vs. sing? (semantic agreement)
    • vote vs. votes? (syntactic agreement)
Constrain Agreements w/ Probe

At the polling station men and women
Constrain Agreements w/ Probe

At the polling station men and women
Plausibility Metric

• Fraction of test cases that pass the test
Define *Trustworthy*.

- Plausible
- **Faithful**
Faithfulness

• The interpretation should be faithful to the model decision mechanism
Challenge

Lack of actionable definition of “model decision mechanism”
Faithfulness (Proxy)

• The interpretation should be faithful to the model decision mechanism resilient to changes that should preserve the model decision mechanism
Recall...

- Saliency is >>>>

- Regarding a prediction $y$ made by a model $M$, give a distribution of feature contributions over the input feature set $F$. 
Input Consistency

- Fix $y$ and $M$, modify $F$ in a way such that the model decision mechanism should be preserved

- Example:
  - The *wife* checked the *son* for injuries because (she...)
  - The *mother* checked the *boy* for injuries because (she...)

Shuoyang Ding — Evaluating Saliency Methods for Neural LM
Model Consistency

- Fix $y$ and $F$, modify $M$ in a way such that the model decision mechanism should be preserved.

- Example:
  - Train a model $M'$ by distilling knowledge from $M$. 

Faithfulness Metric

- Pearson correlation between the feature importance distributions before/after the modification
Data Construction

- Extract prefixes preceding verbs/pronouns
- Filtered to cover cases:
  - with both cues and attractors for plausibility test
  - that are not too trivial (e.g. verb immediately follows the grammatical subject)

<table>
<thead>
<tr>
<th></th>
<th>natural</th>
<th>synthetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>number</td>
<td>PTB</td>
<td>Syneval</td>
</tr>
<tr>
<td>gender</td>
<td>CoNLL</td>
<td>Winobias</td>
</tr>
</tbody>
</table>
## Data Construction

<table>
<thead>
<tr>
<th>PTB</th>
<th>U.S. Trade Representative Carla Hills said the first dispute-settlement panel set up under the U.S.-Canadian “free trade” agreement has ruled that Canada’s restrictions on exports of Pacific salmon and herring (PLURAL…)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syneval</td>
<td>the consultant that loves the parents (SINGULAR…)</td>
</tr>
<tr>
<td>CoNLL</td>
<td>Israeli Prime Minister Ehud Barak says he is freezing tens of millions of dollars in tax payments to the Palestinian Authority. Mr. Barak says he is withholding the money until the Palestinians abide by cease-fire agreements. Earlier Thursday Mr. Barak ruled out an early resumption of peace talks, even with the United States acting as intermediary. Eve Conette reports from Jerusalem. Defending what (MASCULINE…)</td>
</tr>
<tr>
<td>Winobias</td>
<td>The bride examined the son for injuries because (FEMININE…)</td>
</tr>
</tbody>
</table>
Experiment

- Architectures:
  - LSTM
  - QRNN model (Bradbury et al. 2017)
  - Transformer model w/ adaptive input (Baevski and Auli, 2018)
- LMs are trained on WikiText-103
- Probe fine-tuning was done on WikiText-2
Experiment

• Interpretation methods tested:
  • Vanilla Gradient (V)
  • SmoothGrad (SG) (Smilkov et al. 2017)
  • Integrated Gradient (IG) (Sudararajan et al. 2018)
Main Res. (Plausibility)

![Graphs showing comparison of different architectures (LSTM, QRNN, Transformer) for various datasets (PTB, CoNLL, Syneval, Winobias) with different saliency methods (V, SG, IG) and their average accuracy.](image-url)
Main Res. (Faithfulness - Input)

Syneval

Winobias
Main Res. (Faithfulness - Model)

PTB

CoNLL

Syneval

Winobias
## Results

- **How do the failure cases look?**

<table>
<thead>
<tr>
<th>Case</th>
<th>Model</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>QRNN+SG</td>
<td>The [grandmother] examined the (grandson) for injuries because</td>
</tr>
<tr>
<td>1b</td>
<td>QRNN+SG</td>
<td>The [sister] examined the (groom) for injuries because</td>
</tr>
<tr>
<td>2a</td>
<td>QRNN+V</td>
<td>The [grandmother] examined the (grandson) for injuries because</td>
</tr>
<tr>
<td>2b</td>
<td>QRNN+V</td>
<td>The [aunt] examined the (groom) for injuries because</td>
</tr>
<tr>
<td>3a</td>
<td>QRNN+SG</td>
<td>The [grandmother] examined the (grandson) for injuries because</td>
</tr>
<tr>
<td>3b</td>
<td>QRNN_distilled+SG</td>
<td>The [grandmother] examined the (grandson) for injuries because</td>
</tr>
<tr>
<td>4a</td>
<td>Transformer+SG</td>
<td>The [grandmother] examined the (grandson) for injuries because</td>
</tr>
<tr>
<td>4b</td>
<td>Transformer+SG</td>
<td>The [aunt] examined the (groom) for injuries because</td>
</tr>
<tr>
<td>4c</td>
<td>Transformer_distilled+SG</td>
<td>The [grandmother] examined the (grandson) for injuries because</td>
</tr>
</tbody>
</table>
Results

• Can we generalize trend across model config?

No.
Results

- How does distilled student model compare to their teachers?

![Graph showing comparison of average accuracy for LSTM, QRNN, and Transformer architectures with different saliency methods. The graph includes bars for Syneval and plausibility measures.]

Syneval, plausibility
Results

• How does distilled student model compare to their teachers?

Winobias, input faithfulness
Conclusion

• We construct a comprehensive benchmark to evaluate saliency methods in the context of language modeling.

• An interpretation is trustworthy only when both plausibility and faithfulness tests are checked.

• The interpretation quality varies, and sensitive to architectures and model configurations.
Takeaway

Evaluate your interpretation quality before drawing conclusions from them!
Thanks!

Code and data available at:
https://github.com/shuoyangd/tarsius

Email: dings@jhu.edu
Twitter: @_sding
GitHub: shuoyangd