Interpretability in NLP: Moving Beyond Vision

Shuoyang Ding

Microsoft Translator Talk Series
Oct 10th, 2019

Work done in collaboration with Philipp Koehn and Hainan Xu
Outline

• A Quick Tour of Interpretability
  • Model Transparency
  • Post-hoc Interpretations

• Moving Visual Interpretability to Language:
  • Word Alignment for NMT Via Model Interpretation
  • Benchmarking Interpretations Via Lexical Agreement

• Future Work
Outline

• A Quick Tour of Interpretability
  • Model Transparency
  • Post-hoc Interpretations

• Moving Visual Interpretability to Language:
  • Word Alignment for NMT Via Model Interpretation
  • Benchmarking Interpretations Via Lexical Agreement

• Future Work
NOT SURE IF I KNOW

WHAT'S INTERPRETABILITY
What is Interpretability?

- No consensus!
- Categorization proposed in [Lipton 2018]
  - Model Transparency
  - Post-hoc Interpretation
Toy Example
Toy Example
A Transparent Model
Transparent Models

• Build another model that accomplishes the same task, but with easily explainable behaviors.

• Deep neural networks are not interpretable...

• So what models are? (Open question)
  • log-linear model?
  • attention model?
Meh. Too lazy for that!
Post-hoc Interpretation

- **Ask a human**
  - Interpretation with stand-alone model (different task!)

- **Jiggle the cable!**
  - Interpretation with sensitivity w.r.t. features
Post-hoc Interpretation

• Ask a human
  • Interpretation with stand-alone model (different task!)

• Jiggle the cable!
  • Interpretation with sensitivity w.r.t. features
A Little Abstraction...
A Little Abstraction...
A Little Abstraction...

Black Box Model

Δx
A Little Abstraction...
Relative Sensitivity...?

\[ \frac{\Delta y}{\Delta x} \]
Relative Sensitivity...

\[ \frac{\Delta y}{\Delta x} \]

when \( \Delta x \rightarrow 0 \):

\[ \frac{\Delta y}{\Delta x} \rightarrow \frac{\partial y}{\partial x} \]
Saliency

\[
\frac{\partial y}{\partial x}
\]
What’s good about this?

1. **Model-agnostic**, and yet with **some exposure** to the interpreted model

2. Derivatives are **easy to obtain** for any DL toolkit
Saliency in Computer Vision

Image

Saliency

https://pair-code.github.io/saliency/
SmoothGrad

- Gradients are very **local** measure of sensitivity.
- Highly non-linear models may have pathological points where the gradients are **noisy**.

[Smilkov et al. 2017]
SmoothGrad
SmoothGrad

- Solution: calculate saliency for multiple copies of the same input corrupted with gaussian noise, and average the saliency of copies.
SmoothGrad
SmoothGrad in Computer Vision

Original Image

Vanilla

SmoothGrad

https://pair-code.github.io/saliency/
Integrated Gradients (IG)

- Proposed to solve **feature saturation**
- **Baseline**: an input that carries no information
- Compute gradients on **interpolated** baseline & input and average by integration

[Sundararajan et al. 2017]
IG in Computer Vision

Original Image

Vanilla

SmoothGrad

Integrated Gradients

https://pair-code.github.io/saliency/
Summary

Model Transparency:
• Build model that operates in an explainable way
• Interpretation does not depend on output

Post-hoc interpretation:
• Keep the original model intact
• Interpretation depends on specific output
Summary

• How is this related to what I’m talking about next?

• Word Alignment for NMT Via Model Interpretation
  • transparent models vs. post-hoc interpretations

• Benchmarking Interpretations Via Lexical Agreement
  • different post-hoc interpretation methods
Outline

• A Quick Tour of Interpretability
  • Model Transparency
  • Post-hoc Interpretations

• Moving Visual Interpretability to Language:
  • Word Alignment for NMT Via Model Interpretation
  • Benchmarking Interpretations Via Lexical Agreement

• Future Work
Word Alignment

We do not believe that we should cherry-pick.

Wir glauben nicht, daß wir nur rosinen herauspicken sollten.
We believe not, that we only raisin pick should.
Word Alignment

We do not believe that we should cherry-pick.

Wir glauben nicht, daß wir nur rosinen herauspicken sollen.
We do not believe that we

Wir glauben nicht, daß wir nur rosinen herauspicken sollten.
Model Transparency?

We do not believe that we

A Great NMT Model

Wir glauben nicht, daß wir nur rosenen herauspicken sollten.

We believe not, that we only raisin pick should.

Wait... word alignments should be aware of the output!
Post-hoc Interpretations with Stand-alone Models?

\[ p(\alpha_{ij} \mid e, f) \]

*Hint: GIZA++, fast-align, etc.*
Post-hoc Interpretations with Perturbation/Sensitivity?

We do not believe that we should

A Great NMT Model

$\Delta x$
Post-hoc Interpretations with Perturbation/Sensitivity?

We do not believe that we
“Feature” in Computer Vision

Photo Credit: Hainan Xu
“Feature” in NLP

It’s straight-forward to compute saliency for a single dimension of the word embedding.
“Feature” in NLP

We do not believe that we should

A Great NMT Model

But how to **compose** the saliency of **each dimension** into the saliency of a **word**?
Li et al. 2016

*Visualizing and Understanding Neural Models in NLP*

\[
\frac{1}{N} \sum_{i=1}^{N} \left| \frac{\partial y}{\partial e_i} \right|
\]

range: \((0, \infty)\)
Our Proposal

Consider word embedding look-up as a **dot product** between the **embedding matrix** and an **one-hot vector**.
Our Proposal

The 1 in the one-hot vector denotes the identity of the input word.
Our Proposal

Let’s perturb that 1 like a real value!
i.e. take gradients with regard to the 1.
Our Proposal

\[
\sum_i e_i \cdot \frac{\partial y}{\partial e_i}
\]

range: \((-\infty, \infty)\)

Recall this is different from Li’s proposal:

\[
\frac{1}{N} \sum_{i=1}^{N} \left| \frac{\partial y}{\partial e_i} \right|
\]
Why is this proposal better?

• A input word may strongly discourage certain translation and still carry a large (negative) gradient.

• Those are salient words, but shouldn’t be aligned.

• Absolute value/L2-norm falls into this pit.
Evaluation

• Evaluation of interpretations is tricky!

• Fortunately, there’s human judgments to rely on.

• Need to do force decoding with NMT model.
Setup

- Architecture: **Convolutional S2S, LSTM, Transformer** (with fairseq default hyper-parameters)

- Dataset: Following Zenkel et al. [2019], which covers **de-en**, **fr-en** and **ro-en**.

- SmoothGrad hyper-parameters: $N=30$ and $\sigma=0.15$
Baselines

- **Attention weights**

- **Smoothed Attention**: forward pass on multiple corrupted input samples, then average the attention weights over samples

- **[Li et al. 2016]**: compute element-wise absolute value of embedding gradients, then average over embedding dimensions

- **[Li et al. 2016] + SmoothGrad**
Convolutional S2S on de-en

![Bar chart showing AER (Alignment Error Rate) for different methods. Methods include Attention, Smoothed Attention, Li+Grad, Li+SmoothGrad, Ours+Grad, Ours+SmoothGrad, fast-align, Zenkel et al. [2019], and GIZA++. The x-axis represents the methods, and the y-axis represents the AER values. The chart shows that Smoothed Attention, Li+SmoothGrad, and Ours+SmoothGrad have the lowest AER values, indicating better performance.]
Attention on de-en

- Better
- Worse

<table>
<thead>
<tr>
<th>Method</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td>35</td>
</tr>
<tr>
<td>LSTM</td>
<td>45</td>
</tr>
<tr>
<td>Transformer</td>
<td>55</td>
</tr>
<tr>
<td>fast-align</td>
<td>25</td>
</tr>
<tr>
<td>Zenkela et al. [2019]</td>
<td>15</td>
</tr>
<tr>
<td>GIZA++</td>
<td>15</td>
</tr>
</tbody>
</table>
Ours+SmoothGrad on de-en

- Better
- Worse

AER

- Conv
- LSTM
- Transformer
- fast-align
- Zenkel et al. [2019]
- GIZA++
Li vs. Ours

(a) Attention

(b) Li

(c) Ours
Li vs. Ours

(English: We do not believe that we should cherry-pick.)
Summary

• For each of these interpretation methods:
  
  • Attention: maximum transparency on **how the model works**, but is hard to **interpret**
  
  • Stand-alone Alignment Models: gives **best word alignments**, but has nothing to do with the **translation model**
  
  • Saliency: **a good combination of both worlds!**
Outline

• A Quick Tour of Interpretability
  • Model Transparency
  • Post-hoc Interpretations

• Moving Visual Interpretability to Language:
  • Word Alignment for NMT Via Model Interpretation
  • Benchmarking Interpretations Via Lexical Agreement

• Future Work
How about other NLP tasks?

- **Text Classification:**
  [Aubakirova and Bansal 2016][Arras et al. 2016]

- **Sentiment Analysis:**
  [Li et al. 2016][Arras et al. 2017]

- **Question Answering:**
  [Mudrakarta et al. 2018]
Assumption

Post-hoc Interpretation

= How did the model make decision
Assumption

Post-hoc Interpretation

How did the model make decision?
Quick Flashback

We do not believe that we should

Wir glauben nicht, daß wir nur rosinen herauspicken sollen.

We believe not, that we only raisin pick should.
Quick Flashback

We do not believe that we should

A Great NMT Model

Wir glauben nicht, daß wir nur rosinen herauspicken sollten.
We believe not, that we only raisin pick should.

Li et al. 2016

We do not believe that we should

A Great NMT Model

Wir glauben nicht, daß wir nur rosinen herauspicken sollten.
We believe not, that we only raisin pick should.

Ours+SmoothGrad
Research Question

• How can we quantitatively test the effectiveness of model interpretation methods in the context of NLP?

• What are the said “effectiveness” correlated with? model size? architecture? task performance?
Computer Vision

Yao et al. 2018
Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions
Main Challenge

No ground-truth interpretation
Lexical Agreements

• Frequently studied for interpretability [Linzen et al. 2016][Marvin and Linzen 2018][Gulordava et al. 2018][Giulianelli et al. 2018]

• They concentrate on evaluating probing task performance, i.e. whether the model can predict the lexical agreements properly
However, most people, having been subjected to news footage of the devastated South Bronx, ...

A. look  B. looks
E.g. Subject-Verb Agreements

However, most *people*, having been subjected to news *footage* of the devastated South *Bronx*, ...

A. look   B. looks
E.g. Subject-Verb Agreements

However, most **people**, having been subjected to news **footage** of the devastated South **Bronx**, ...

**A. look**
E.g. Subject-Verb Agreements

However, most people, having been subjected to news footage of the devastated South Bronx, ...

A. look   B. looks

“Probing Task”
The Test

However, most people, having been subjected to news footage of the devastated South Bronx, look
However, most people, having been subjected to news footage of the devastated South Bronx, looks
The Test

However, most people, having been subjected to news footage of the devastated South Bronx, look

The interpretation passes the test, if $\forall w \in \{\text{footage, Bronx}\}$, s.t.

$$\psi(\text{people}) > \psi(w)$$

$\psi$: feature importance/saliency
The Test

However, most people, having been subjected to news footage of the devastated South Bronx, looks

The interpretation passes the test, if \( \exists w \in \{\text{footage, Bronx}\} \), s.t.

\[
\psi(\text{people}) < \psi(w)
\]

\( \psi \): feature importance/saliency
The Test

- We constructed test set based on two existing human-annotated corpus
  - **Penn Treebank**: new, multiple attractors
  - **syneval**: Marvin and Linzen [2018], single attractor
- We plan to construct another one with **CoNLL-2012 coreference resolution dataset** -- stay tuned!
Interpreted Model

• Language Model!

• With final linear layer replaced with one that is fine-tuned for predicting specific agreement of interest

• Word prediction may introduce out-of-scope agreements and interfere with evaluation
Experiment

- Architectures:
  - **LSTM model**, trained on WikiText-2
  - **QRNN model** [Bradbury et al. 2017], trained on WikiText-2
  - **Transformer model w/ adaptive input** [Baevski and Auli, 2018], trained on WikiText-103
  - All the fine-tuning was done on WikiText-2
  - For subject-verb agreement, the verb tagging is done with Stanford POS-tagger
Probing Task Performance

![Bar chart showing performance comparison between LSTM, QRNN, and Transformer for "penn" and "syneval" tasks. The bars indicate a higher performance for LSTM and Transformer compared to QRNN.](image)
Interpretation of LSTM

- random
- vanilla
- li
- li_smoothed
- smoothed
- integral

Graph showing performance across different datasets (penn, syneval) with various model modifications.
Interpretation of QRNN

- penn
- syneval

- random
- vanilla
- li
- li_smoothed
- smoothed
- integral

Graph showing performance metrics for different models and datasets.
Interpretation of Transformer

![Chart showing the interpretation of Transformer with different methods: random, vanilla, li, li_smoothed, and smoothed. The chart compares penn and syneval.](image-url)
What's up with Transformer?

- Two hypothesis:
  - **Deep model** hurts interpretability
  - **Too many heads** hurts interpretability

- SOTA model: 16 layers, 8 heads

- Diagnostic model:
  - 4 layers, 8 heads
  - 4 layers, 1 head
16 layers, 8 heads

- random
- vanilla
- li
- li_smoothed
- smoothed
- integral

penn  syneval
4 layers, 8 heads

- random
- li
- smoothed
- vanilla
- li_smoothed
- integral

Bar chart comparing different models on 'penn' and 'syneval' datasets.
4 layers, 4 heads

- random
- li
- li_smoothed
- smoothed
- integral

Comparison between penn and syneval.
4 layers, 2 heads

- random
- vanilla
- li
- li_smoothed
- smoothed
- integral

- penn
- syneval
Some Qualitative Checks

- Are those interpretations just looking at the immediate previous word?
- No. They seems to get a lot of things right!
Some Qualitative Checks

- Are they the same with different architectures?

  - No. Different architectures work differently.
Summary

• Lexical agreements open up possibilities to do **rigorous quantitative checks** for post-hoc interpretation methods in the context of NLP

• Our proposed method **works the best** consistently

• **Deep NLP models** can be **out-of-reach** for existing interpretation methods.
Outline

• A Quick Tour of Interpretability
  • Model Transparency
  • Post-hoc Interpretations

• Moving Visual Interpretability to Language:
  • Word Alignment for NMT Via Model Interpretation
  • Benchmarking Interpretations Via Lexical Agreement

• Future Work
Future Work

- **Better interpretation method** that works for the deep architectures in NLP.

- How can we use interpretability in **real-world applications (QE?)**, or **improve our models**?

- How can we use interpretability to validate whether the model learned certain **linguistic properties**?
Thanks!

**email**: dings@jhu.edu  
**twitter**: @_sding  
**github**: shuoyangd