Analysis methods in NLP: Probing

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding
1. Core idea: use supervised models (the probes) to determine what is latently encoded in the hidden representations of our target models.

2. Often applied in the context of BERTology – see especially Tenney et al. 2019.

3. A source of valuable insights, but we need to proceed with caution:
   - A very powerful probe might lead you to see things that aren’t in the target model (but rather in your probe).
   - Probes cannot tell us about whether the information that we identify has any causal relationship with the target model’s behavior.

Core method

Conneau et al. 2018; Tenney et al. 2019
### Core method

<table>
<thead>
<tr>
<th>a</th>
<th>c</th>
<th>f</th>
<th>m</th>
<th>r</th>
<th>w</th>
<th>t</th>
</tr>
</thead>
</table>

$\text{SmallLinearModel}(h) = \text{task}_X \cdot \text{task}_y$

$h_1 \cdot \text{task}_1 \cdot \text{task}_y$

$h_2 \cdot \text{task}_2 \cdot \text{task}_y$

$h_3 \cdot \text{task}_3 \cdot \text{task}_y$

---

Conneau et al. 2018; Tenney et al. 2019
Core method

SmallLinearModel(h) = task

Conneau et al. 2018; Tenney et al. 2019
Core method

Conneau et al. 2018; Tenney et al. 2019
Core method

\[
\text{SmallLinearModel}(h) = \text{task}_1 \quad y
\]

Conneau et al. 2018; Tenney et al. 2019
Core method

Conneau et al. 2018; Tenney et al. 2019
Core method

SmallLinearModel($h$) = task

Conneau et al. 2018; Tenney et al. 2019
Core method

SmallLinearModel\( (h) = \text{task}\ X y \text{task} 1 y \text{task} 2 y \text{task} 3 \)

Conneau et al. 2018; Tenney et al. 2019
Probing or learning a new model?

1. Probes in the above sense are supervised models whose inputs are frozen parameters of the model we are probing.

2. This is hard to distinguish from simply fitting a supervised model as usual, with a particular choice for featurization.

3. At least some of the information that we identify is likely to be stored in the probe model.

4. More powerful probes might “find” more information – by storing more information in the probe parameters.
Control tasks and probe selectivity

Control task
A random task with the same input/output structure as the target task.

- Word-sense classification: words assigned random fixed senses.
- POS tagging task: words assigned random fixed tags.
- Parsing: assigned edges randomly using simple strategies.

Selectivity
The difference between probe performance on the task and probe performance on the control task.

Hewitt and Liang 2019
Control tasks and probe selectivity

![Graph showing the relationship between accuracy and selectivity with increasing complexity]

- High Accuracy, High Selectivity
- High Accuracy, Low Selectivity

**Legend:**
- Part-of-speech task
- Control task
- Selectivity

**Hewitt and Liang 2019**
A fundamental limitation: No causal inference

Belinkov and Glass 2019; Vig et al. 2020
A fundamental limitation: No causal inference

Belinkov and Glass 2019; Vig et al. 2020
A fundamental limitation: No causal inference

1. Probe $L_1$: it computes $x + y$

Belinkov and Glass 2019; Vig et al. 2020
A fundamental limitation: No causal inference

1. Probe $L_1$: it computes $x + y$
2. Probe $L_2$: it computes $z$

Belinkov and Glass 2019; Vig et al. 2020
A fundamental limitation: No causal inference

1. Probe $L_1$: it computes $x + y$
2. Probe $L_2$: it computes $z$
3. But neither has any impact on the output!

$$W_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad W_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad W_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$w = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Model:

$$(xW_1; xW_2; xW_3) w$$

Belinkov and Glass 2019; Vig et al. 2020
Unsupervised probes

2. Clark et al. (2019) and Manning et al. (2020): Inspecting attention weights.
3. Hewitt and Manning (2019) nd Chi et al. (2020): Linear transformations of hidden states to identify latent syntactic structures in BERT.
4. Rogers et al. (2020): extensive discussion of probing and related efforts and what they have revealed about BERT representations.


