Visualizing and Understanding Neural Machine Translation

Yanzhuo Ding, Yang Liu, Huanbo Luan, Maosong Sun
Tsinghua University

Presented by: Yuchen He
source words
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
source contexts
target hidden states
target word embeddings
target words
Layer-wise relevance propagation (LRP)

Can calculate the relevance between two arbitrary neurons
Measures/visualizes how much each pixel is related to the final classification
Goal

● To quantify and visualize the relevance between a neural network layer and contextual word vectors (source & target word embeddings)

Offers more insights in interpreting how target words are generated
Relevance vector

source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

source contexts

target hidden states

target word embeddings

target words
Calculating Neuron-Level Relevance

Base case: (relevance of \( v \) to itself)

\[
\mathcal{R}_{v \leftarrow v} = \mathcal{U} \quad \text{for any neuron } v
\]

Recursive case: (relevance of \( u \) to \( v \))

\[
\mathcal{R}_{u \leftarrow v} = \sum_{z \in \text{OUT}(u)} \mathcal{W}_{u \rightarrow z} \mathcal{R}_{z \leftarrow v} \quad \text{for any neurons } u, v
\]

\( \text{OUT}(u) \) comprises all \( u \)'s \textit{directly connected descendant} neurons in the network.
Calculating Weight Ratios

\[ w_{u \rightarrow v} = \frac{W_{u, v}}{\sum_{u' \in \text{IN}(v)} W_{u', v}} \]

for any neurons \( u, v \)

\( W_{u, v} \) is the weight of \( u \) to \( v \) in the existing neural network \( \text{IN}(u) \) comprises all \( u \)'s \textit{directly connected ancestor} neurons in the network.
Putting things together

Sum up \( r_{u_n \leftarrow v_m} \) and get vector-level relevance

\[
R_{u \leftarrow v} = \sum_{m=1}^{M} \sum_{n=1}^{N} r_{u_n \leftarrow v_m}
\]

Generate and normalize relevance vector \( R_v \) as a sequence of \( R_{u \leftarrow v} \) for all related contextual word vectors

\[
R_v = \{ R_{u_1 \leftarrow v}, \ldots, R_{u_{|c(v)|} \leftarrow v} \}
\]
Relevance vector

source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

source contexts

target hidden states

target word embeddings

target words
Application

Help debug attention-based NMT systems

- Word omission
- Word repetition
- Unrelated words
- Negation reversion
“Relevance matrix”

- Attention weights
- Source context vector
- Target hidden state
- Target word embedding

Graphical representation of relevance matrix with the words "visit", "is", and "pray".
“Relevance matrix”

**attention weights**

**source context vector**

**target hidden state**

**target word embedding**

<table>
<thead>
<tr>
<th></th>
<th>visit</th>
<th>is</th>
<th>to</th>
<th>pray</th>
</tr>
</thead>
<tbody>
<tr>
<td>my</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>我</td>
<td>参拜</td>
<td>是</td>
<td>为了</td>
<td>祈求</td>
</tr>
<tr>
<td>Input</td>
<td>巴基斯坦总统穆沙拉夫赢得参众两院信任投票</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Pakistani president Musharraf wins votes of confidence in senate and house</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Output</td>
<td>Pakistani president win over democratic vote of confidence in senate (missing words)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
vote of confidence in the senate confidence vote
In the history of the history of the history of the history of the Americans, there is a tradition of faith in the history of mistakes.
Americans' history have
Americans have history.
<table>
<thead>
<tr>
<th>Input</th>
<th>此次会议的一个重要议题是跨大西洋关系</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>One of the top agendas of the meeting is to <strong>discuss</strong> the transatlantic relations</td>
</tr>
<tr>
<td>Output</td>
<td>A key topic of the meeting is to <strong>forge ahead</strong></td>
</tr>
</tbody>
</table>
is transatlantic relations
is transatlantic relations
<table>
<thead>
<tr>
<th>Negation reversion</th>
</tr>
</thead>
</table>

| Input             | 不解决生存问题，就谈不上发展，更谈不上可持续发展 |
| Reference         | Without solving the issue of subsistence, there will be no development to speak of, let alone sustainable development |
| Output            | If we do not solve the problem of living, we will talk about development and still less can we talk about sustainable development |
we_11 will_12 talk_13 about_14 development_15

\[ \alpha_{14} \]
\[ R_{c_{14}} \]
\[ R_{s_{14}} \]
\[ R_{y_{14}} \]

就_6 谈_7 不_8 上_9 发展_10 will_12 talk_13

talke not
development
we \textsubscript{11} \hspace{1em} will \textsubscript{12} \hspace{1em} talk \textsubscript{13} \hspace{1em} about \textsubscript{14} development \textsubscript{15}
Thank you