Visualizing and Understanding Neural Machine Translation

Yanzhuo Ding, Yang Liu, Huanbo Luan, Maosong Sun

Tsinghua University
Neural Machine Translation

- Idea: using neural networks to translate languages (Bahdanau et al., 2015)

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

我 喜欢 温哥华 ＜/s＞

I like Vancouver ＜/s＞

attention
Challenge

• It is hard to visualize and understand the internal workings

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
Related Work

attention mechanism
(Bahdanau et al., 2015)

restricted to the connection between input and output

first-derivative saliency
(Li et al., 2016)

require neural activations to be differentiable
Related Work

layer-wise relevance propagation (LRP)  
(Bach et al., 2015)

calculating the relevance between two arbitrary neurons without requiring differentiability
Our Work

source words

我 喜欢 温哥华 </s>
Our Work

source words

source word embeddings
Our Work

source words

source word embeddings

source forward hidden states
Our Work

source words

source word embeddings

source forward hidden states

1.00

我 喜欢 温哥华 </s>

targeted vector of neurons
relevant vector of neurons
intermediate vector of neurons
irrelevant vector of neurons
1.0 relevance
Our Work

source words

source word embeddings

source forward hidden states
Our Work

source words

source word embeddings

source forward hidden states
Our Work

source words

source word embeddings

source forward hidden states

我 喜欢 温哥华 </s>
Our Work

source words

source word embeddings

source forward hidden states

source backward hidden states

我 喜欢 温哥华 </s>
source words
source word embeddings
source forward hidden states
source backward hidden states
Our Work

source words

source word embeddings

source forward hidden states

source backward hidden states
Our Work

source words

source word embeddings

source forward hidden states

source backward hidden states

我 喜欢 温哥华 </s>
source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states
Our Work

0.12 0.72 0.11 0.05

source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states
Our Work

source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

我 喜欢 温哥华 </s>
source words
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
Our Work

source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

source contexts

0.58
0.32
0.07
0.03

我 喜欢 温哥华 </s>

attention
source words
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
source contexts
target hidden states

Our Work

我 喜欢 温哥华 </s>
Our Work

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

0.78 0.12 0.07 0.03

我 喜欢 温哥华 </s>

attention
Our Work

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

我 喜欢 温哥华 </s>

0.18 0.29 0.25 0.03

attention

0.23
Our Work

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

我 喜欢 温哥华 </s>

attention

0.14 0.33 0.24 0.04
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

0.11 0.45 0.11 0.02

attention

像 喜欢 温哥华 </s>

我

像

like
Our Work

- 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
Our Work

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

我 喜欢 温哥华 </s>

attention

0.05 0.16 0.35 0.02

0.11 0.31
Our Work

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
Our Work

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
Our Work

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

\[
\begin{align*}
\text{我} & \quad \text{喜欢} & \quad \text{温哥华} & \quad \text{}</s>
\end{align*}
\]

attention

\[
\begin{align*}
0.02 & \quad 0.06 & \quad 0.15 & \quad 0.51
\end{align*}
\]
Our Work

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

attention

我 喜欢 温哥华</s>
Vector- and Neuron-Level Relevance

vector-level relevance  neural-level relevance
Neuron-Level Relevance

• Idea: decompose the activation of the targeted neuron among relevant neurons

\[
\mathbf{u} = u_1^3 \quad \mathbf{z} = z_1^3 \quad \mathbf{v} = v_1^3
\]

\[
\mathbf{v}_m = \sum_{\mathbf{u} \in C(v_m)} \sum_{n=1}^{N} r_{u_n} {\leftarrow} v_m
\]

For example

\[
\mathbf{v}_1 = \sum_{n=1}^{3} r_{u_n} {\leftarrow} v_1
\]
Calculating Neuron-Level Relevance

• Recursive calculation in a backward propagation

\[
\begin{align*}
\mathbf{u} &= u_1^3 \\
\mathbf{z} &= z_1^3 \\
\mathbf{v} &= v_1^3
\end{align*}
\]

\[
\mathbf{r}_{u \leftarrow v} = \sum_{z \in \text{OUT}(u)} w_{u \rightarrow z} \mathbf{r}_{z \leftarrow v}
\]

For example

\[
\begin{align*}
\mathbf{r}_{u_1 \leftarrow v_1} &= \sum_{k=1}^{3} w_{u_1 \rightarrow z_k} \mathbf{r}_{z_k \leftarrow v_1} \\
\mathbf{r}_{z_k \leftarrow w_1} &= w_{z_k \rightarrow v_1} v_1
\end{align*}
\]
Calculating Weight Ratios

• Recursive calculation in a forward propagation

\[
\begin{align*}
\mathbf{u} &= u_1^3 \\
\mathbf{z} &= z_1^3 \\
\mathbf{v} &= v_1^3
\end{align*}
\]

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

For example

\[
\begin{align*}
w_{u_1 \rightarrow z_1} &= \frac{W_{1,1}^{(1)} u_1}{W_{1,1}^{(1)} u_1 + W_{2,1}^{(1)} u_2 + W_{3,1}^{(1)} u_3}
\end{align*}
\]
Calculating Vector-Level Relevance

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

- Specify targeted vector of neurons
- Calculate weight ratios in a forward propagation
- Calculate relevance in a backward propagation

*layer-wise propagation for neural machine translation*
## Application

- Help to debug attention-based NMT systems

<table>
<thead>
<tr>
<th>word omission</th>
<th>word repetition</th>
<th>unrelated words</th>
<th>negation reversion</th>
</tr>
</thead>
</table>

Analyzing major translation error types by visualizing relevance step by step
Open-Source Toolkit

http://thumt.thunlp.org
• It is challenging to interpret how neural networks work
• We leverage layer-wise relevance propagation to visualize NMT
• Our approach can be applied to networks in other NLP tasks
Thanks

http://thumt.thunlp.org