Recurrent Neural Network Attention Mechanisms for Interpretable System Log Anomaly Detection

Presented By:
Akash Kulkarni
System Log Analysis is complicated

1. Log sources generate TBs of data per day
2. Lack of labelled data (scarce or unbalanced)
3. Actionable information may be obscured (by complex relationships across logging sources)

Need for an aided human monitoring and assessment.
Unsupervised RNN language models

• Models distribution of normal events in system logs (learns complex relationships buried in logs)
• No need of labelled data.
• No feature engineering required (as deep learning learns significant features automatically)
Language Modelling

• Each log-line consists of sequence of T tokens:
  • \( x_{(1:T)} = x_{(1)}, x_{(2)}, x_{(3)} \ldots \ldots x_{(T)} \), each token \( x_{(i)} \in V \)

• Language model (like RNNs) assigns probabilities to sequences:
  • \( P(x_{(1:T)}) = \prod_{t=1}^{T} P(x_{(t)} | x_{(<t)}) \)

• Tokenization can be word-based or character-based.
Cyber Anomaly Language Models

1. Event Model (EM)—which applies a standard LSTM to log lines

\[
\text{LSTM}(\mathbf{x}_{(1:T)}) = \mathbf{h}_{(1:T)}
\]

\[
p(t) = \text{softmax}\left(\frac{\mathbf{h}_{(t-1)} \mathbf{W} + \mathbf{b}}{L_h} \right) \in \mathbb{R}^{|V|}
\]

2. Bidirectional Event Model (BEM)

\[
p(t) = \text{softmax}\left(\frac{\mathbf{h}_{(t-1)} \mathbf{W} + \mathbf{h}_{(t+1)}^b \mathbf{W}^b + \mathbf{b}}{L_h \times |V| + |V|} \right)
\]
Cyber Anomaly Language Models

3. Tiered Language Model (T-EM or T-BEM)
Attention

- Key matrix, $K = \tanh(VW^a)$
- Weights, $d = \text{softmax}(qK^T)$
- Attention, $a = dV$

• Prediction, $p(t) = \text{softmax} \left( \begin{bmatrix} h_{t-1} & a_{t-1} \end{bmatrix} W + b \right)$
EM attention variants

1. Fixed Attention:
   • $q(t) = q$
   • Assumes some position in the sequence are more important than others

2. Syntax Attention:
   • $q(t)$ not shared across $t$
   • Importance depends on the position of the current token in sequence.

3. Semantic Attention 1:
   • $q(t) = \tanh(h(t)W^{sem1})$

4. Semantic Attention 2:
   • $h'(t) = \text{concatentation of } h(t) \text{ and } q(t)$
5. Tiered Attention
   • Replaces mean with weighted average via attention

\[
q = \tanh(h(T)W^{(tier)})
\]
\[
K = \tanh(VW^a)
\]
\[
d = \text{softmax}(qK^T)
\]
\[
a = dV
\]
## Results

### Table 1. AUC statistics for word tokenization models

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>0.968</td>
<td>0.976</td>
<td>0.964</td>
<td>0.005</td>
</tr>
<tr>
<td>BEM</td>
<td>0.976</td>
<td>0.981</td>
<td>0.972</td>
<td>0.003</td>
</tr>
<tr>
<td><strong>EM with attention</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>0.974</td>
<td>0.976</td>
<td>0.972</td>
<td>0.001</td>
</tr>
<tr>
<td>Syntactic</td>
<td>0.972</td>
<td>0.975</td>
<td>0.967</td>
<td>0.004</td>
</tr>
<tr>
<td>Semantic 1</td>
<td>0.975</td>
<td>0.980</td>
<td>0.971</td>
<td>0.004</td>
</tr>
<tr>
<td>Semantic 2</td>
<td>0.973</td>
<td>0.976</td>
<td>0.968</td>
<td>0.003</td>
</tr>
<tr>
<td><strong>Tiered LSTM variants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-EM</td>
<td>0.984</td>
<td>0.989</td>
<td>0.977</td>
<td>0.005</td>
</tr>
<tr>
<td>T-BEM</td>
<td>0.987</td>
<td>0.989</td>
<td>0.985</td>
<td>0.002</td>
</tr>
<tr>
<td>TA-EM</td>
<td>0.985</td>
<td>0.991</td>
<td>0.979</td>
<td>0.004</td>
</tr>
<tr>
<td>TA-BEM</td>
<td>0.988</td>
<td>0.991</td>
<td>0.984</td>
<td>0.003</td>
</tr>
</tbody>
</table>

### Table 2. AUC statistics for character tokenization models

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>0.965</td>
<td>0.969</td>
<td>0.961</td>
<td>0.003</td>
</tr>
<tr>
<td>BEM</td>
<td>0.985</td>
<td>0.987</td>
<td>0.982</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>EM with attention</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>0.963</td>
<td>0.971</td>
<td>0.937</td>
<td>0.015</td>
</tr>
<tr>
<td>Syntactic</td>
<td>0.967</td>
<td>0.973</td>
<td>0.963</td>
<td>0.004</td>
</tr>
<tr>
<td>Semantic 1</td>
<td>0.975</td>
<td>0.977</td>
<td>0.971</td>
<td>0.003</td>
</tr>
<tr>
<td>Semantic 2</td>
<td>0.972</td>
<td>0.977</td>
<td>0.967</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>Tiered LSTM variants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-EM</td>
<td>0.977</td>
<td>0.988</td>
<td>0.967</td>
<td>0.008</td>
</tr>
<tr>
<td>T-BEM</td>
<td>0.992</td>
<td>0.992</td>
<td>0.991</td>
<td>0.000</td>
</tr>
<tr>
<td>TA-EM</td>
<td>0.982</td>
<td>0.984</td>
<td>0.979</td>
<td>0.002</td>
</tr>
<tr>
<td>TA-BEM</td>
<td>0.991</td>
<td>0.992</td>
<td>0.990</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Analysis

Comparison of attention weights when predicting success/failure token
Analysis

1. Average Fixed attention weights

2. Average Syntax attention weights

3. Average Semantic 1 attention weights

4. Average Semantic 2 attention weights
Analysis

• Tiered attention models
  • For lower forward-directional LSTM, attention weights were nearly 1.0 for 2\textsuperscript{nd} to last hidden state.
  • For lower bidirectional LSTM, attention weights were nearly 1.0 for 1\textsuperscript{st} hidden state and last state
  • Hence, attentions are not needed for this model task.
Case Studies

1. Word case study with Semantic attention

2. Low anomaly word case study with Semantic attention

3. Case character study with Semantic attention
Conclusions

• Fixed and syntactic attention effective for fixed structure sequences.
• Attention mechanism improve performance and provide feature importance and relational mapping between features.

Future directions

• Explore BEM with attention
• Equipping a lower tier model to attend over upper tier hidden states