**Introduction**

- Multi-task learning for DM/PSD/UCCA: Propose a unified neural model for DM/PSD/UCCA frameworks based on biaffine attention used in (Dozat and Manning, 2017, 2018; Zhang et al., 2019)

**Model architecture**

- BERT-BiLSTM sentence encoder (shared across frameworks): an input sentence is fed to a word representation layer using BERT, resulting in a sequence of word embedding vectors, which are then given to the BiLSTM layer to produce a sentence representation.

- Biaffine attention decoder (framework-specific): additional feed-forward layers are applied to obtain role-dependent representations for head and dependent roles, which are then forwarded to the biaffine attention.

**Decoder: Biaffine attention**

- Biaffine attention decoder: perform the role-dependent representations to predict the existence of an edge and its labels.

\[
Biaff_{m}(x, y) = X^{T} \gamma_{[m]} y + V \begin{bmatrix} x \ y \end{bmatrix} + b
\]

- The top score \( S_{\text{top}}(i) \) is newly introduced in our model.

**Multi-level biaffine attention**

- Motivation: Predicting an arc and a label may be resolved not just by single-level representation but by the combination of various levels of representations.

**Word representation layer using BERT:** Given a sentence, the BERT encoder is applied to its wordpieces and the encoded wordpiece-level representations are composed to the word-level embeddings based on BiLSTM.

**Encoder: BERT-BiLSTM**

- Use a simple BiLSTM with a single output layer

\[
x_{i}^{(\text{prop})} = BiLSTM_{i}(x_{1} \cdots x_{n})
\]

**Preliminary Experiment**

<table>
<thead>
<tr>
<th>method</th>
<th>DM</th>
<th>UF</th>
<th>LF</th>
<th>PSD</th>
<th>UF</th>
<th>LF</th>
<th>UCCA</th>
<th>UF</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaffine</td>
<td>93.67</td>
<td>92.08</td>
<td>90.86</td>
<td>95.97</td>
<td>90.50</td>
<td>80.21</td>
<td>72.60</td>
<td>69.67</td>
<td>65.17</td>
</tr>
<tr>
<td>BERT+Biaffine</td>
<td>95.06</td>
<td>93.85</td>
<td>93.00</td>
<td>98.69</td>
<td>92.90</td>
<td>80.24</td>
<td>77.09</td>
<td>74.85</td>
<td>70.15</td>
</tr>
<tr>
<td>BERT+Multi-level Biaffine</td>
<td>95.09</td>
<td>93.86</td>
<td>93.02</td>
<td>96.76</td>
<td>91.95</td>
<td>79.76</td>
<td>78.12</td>
<td>74.42</td>
<td>69.81</td>
</tr>
<tr>
<td>BERT+BiLSTM+MTL</td>
<td>N/A</td>
<td>93.66</td>
<td>92.73</td>
<td>N/A</td>
<td>92.13</td>
<td>79.63</td>
<td>N/A</td>
<td>75.40</td>
<td>70.59</td>
</tr>
</tbody>
</table>

- Lesson: instead of naively using the shared encoder only, other advanced multi-task learning approaches such as placing task-specific encoding, as detailed in (Peng et al., 2017), need to be considered.