Attention-Focused Adversarial Training for Robust Temporal Reasoning

Lis Kanashiro Pereira¹, Kevin Duh², Fei Cheng³, Masayuki Asahara⁴, Ichiro Kobayashi¹
Ochanomizu University¹, Johns Hopkins University², Kyoto University³, NINJAL⁴

Overview

Goal: Improve model generalization and robustness for NLU

Adversarial Training

\[
\min_{\delta} \mathbb{E}_{(x,y) \sim D}[\max_{\delta} \ell(f(x + \delta; \theta), y) + \alpha \max_{\delta} \ell(f(x + \delta_2; \theta), f(x; \theta))]
\]

- Better generalization and robustness
- In NLP: perturbations are usually added to the embeddings only
- Perturbations are usually generated by running a fixed number of gradient steps

Weaknesses:

- Adding the perturbation to the embeddings only might not be optimal
- Other layers of the transformer based models can encode specific syntactic and semantic information

Solution:

- Add the noise to a combination of the model layers instead of only to the embedding layer
- Adds the adversarial perturbation to multiple hidden states or attention representations of the model layer

ML-ALICE

We fine-tune RoBERTa_BASE on the CosmosQA training data and evaluate it on various Temporal test sets

<table>
<thead>
<tr>
<th>Methods</th>
<th>TEA Acc</th>
<th>TimeML Acc</th>
<th>MC-TACO EM</th>
<th>SCT F1</th>
<th>MATRES F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>95.20</td>
<td>89.40</td>
<td>80.86</td>
<td>68.63</td>
<td>92.95</td>
</tr>
<tr>
<td>FreeLB</td>
<td>95.75</td>
<td>90.62</td>
<td>82.75</td>
<td>71.52</td>
<td>92.68</td>
</tr>
<tr>
<td>SMART</td>
<td>95.32</td>
<td>89.77</td>
<td>82.25</td>
<td>73.07</td>
<td>92.89</td>
</tr>
<tr>
<td>ALICE</td>
<td>95.63</td>
<td>90.35</td>
<td>82.35</td>
<td>73.04</td>
<td>93.43</td>
</tr>
<tr>
<td>ML-ALICE (hidden)</td>
<td>95.69</td>
<td>90.52</td>
<td>83.35</td>
<td>49.25</td>
<td>74.78</td>
</tr>
<tr>
<td>ML-ALICE (attendtion)</td>
<td>96.72</td>
<td>92.80</td>
<td>83.94</td>
<td>74.78</td>
<td>94.07</td>
</tr>
</tbody>
</table>

ML-ALICE is effective in enhancing model generalizability in out domains.

Zero-Shot Results

Default text encoder: RoBERTa_BASE

ML-ALICE with adversarial perturbations added to attention representations performs best among all models.