Large-Scale Adversarial Training for Vision-and-Language Representation Learning

Zhe Gan\textsuperscript{1}, Yen-Chun Chen\textsuperscript{1}, Linjie Li\textsuperscript{1}, Chen Zhu\textsuperscript{2}, Yu Cheng\textsuperscript{1}, Jingjing Liu\textsuperscript{1}

\textsuperscript{1}Microsoft Dynamics 365 AI Research, \textsuperscript{2}University of Maryland, College Park

Motivation & Contribution

- **Multimodal pre-training**, such as ViLBERT, LXMERT and UNITER, has made tremendous progress in Vision-and-Language (V+L) research.
- However, aggressive finetuning of pre-trained models often falls into the overfitting trap.
- **Adversarial training** has shown great potential in improving the generalization ability of BERT for language understanding tasks.
- **Our Contribution**: the first known effort to study large-scale adversarial training for V+L tasks.

Algorithm and Backbone (UNITER)

Experimental Results

<table>
<thead>
<tr>
<th>Task</th>
<th>VQA</th>
<th>VCR</th>
<th>NLVR2</th>
<th>VE</th>
<th>RefCOCOg</th>
<th>RefCOCO+</th>
<th>Flickr30k IR</th>
<th>Flickr30k TR</th>
<th>VQA-Rep.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNITER</td>
<td>74.02</td>
<td>62.8</td>
<td>79.98</td>
<td>79.38</td>
<td>75.77</td>
<td>66.70</td>
<td>75.56</td>
<td>87.30</td>
<td>64.56</td>
</tr>
<tr>
<td>VILLA</td>
<td>74.87</td>
<td>65.7</td>
<td>81.47</td>
<td>80.02</td>
<td>76.71</td>
<td>66.84</td>
<td>76.26</td>
<td>87.90</td>
<td>65.35</td>
</tr>
</tbody>
</table>

Both adversarial pre-training (VILLA-pre) and finetuning (VILLA-fine) contribute to performance boost.

Adversarial training on image or text modality alone is already effective.

VILLA captures richer visual coreference and visual relation knowledge than UNITER.

VILLA learns more accurate and sharper attention maps than UNITER.

VILLA is more robust to paraphrases than UNITER.

Code is available at [https://github.com/zhegan27/VILLA](https://github.com/zhegan27/VILLA)