Semi-supervised Grounding Alignment for Multi-modal Feature Learning

Shih-Han Chou\textsuperscript{1,2}, Zicong Fan\textsuperscript{3}, James J. Little\textsuperscript{1}, Leonid Sigal\textsuperscript{1,2,4}

\textsuperscript{1}Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada
\textsuperscript{2}Vector Institute for AI, Toronto, Ontario, Canada
\textsuperscript{3}Department of Computer Science, ETH Zürich, Rämistrasse, Zürich, Switzerland
\textsuperscript{4}Canada CIFAR AI Chair
{shchou75, little, lsigal}@cs.ubc.ca; zicong.fan@inf.ethz.ch

In this supplementary material, we present additional results to complement the main paper.

I. GROUNDING ALIGNMENT

We show examples of the pseudo ground truth of grounding alignment on the Conceptual Captions dataset in Figure 1.

II. DETAILS OF THE GROUNDING ALIGNMENT MODULE

Figure 2 summarizes the details of the grounding alignment module.

III. HYPER-PARAMETER TUNING

We show hyper-parameter tuning in Table I. The settings with SPE or at phrase level achieve higher accuracy than the settings at token level which indicates that SPE and phrase level grounding are more important.

IV. ADJUSTING LEARNING RATES

The adjustment of learning rates is shown in Table II. We train our model with 0.25x, 0.5x, and 2x of the original setting.

V. ADDITIONAL QUALITATIVE RESULTS

A. Visual Grounding

We show some qualitative results of the visual grounding task in Figure 3. The orange boxes denote the ground truth annotations. The red boxes are the baseline [1] and the blue boxes are ours. The top two rows are the results where we perform better than the baseline. As shown in the first example, the ‘silver truck’, our predicted region is more aligned with the ground truth than the baseline. We also show some failure cases in the bottom row. Although our method does not exactly match the ground truth, the predictions are more reasonable than the baseline, we predict the oven but not the table in the first example in the bottom row.

We also show some examples in Figure 4 for the model pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 RefCOCO+ dataset.

B. Visual Question Answering

We show some qualitative results of the VQA task in Figure 5. The green color means the prediction is the same as the ground truth and the red color represents the wrong answer. The top row shows the positive examples where our method predicts the correct answers but the baseline doesn’t. We also show some failure cases in the bottom row. In the failure case, ‘How many vegetables are there?’, our model is confused by the beans so that the prediction is ‘3’ but not ‘2’. Similarly, in the last example, the color of the curtains is too dark to distinguish whether they are blue or black. We also show some examples in Figure 6 for the model pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 VQA dataset.

C. Visual Commonsense Reasoning

We show some qualitative results of the VCR task in Figure 7. The green color denotes the correct answers and the red color denotes the wrong answers. We show the probabilities among the answer choices. The examples show that our method is able to put the right attention on the correct answers. For the last example, although the answer is (b) in the QA \rightarrow R task, we argue that the answer choices are somewhat ambiguous. Answers (a) and (b) are both suitable for this question. We also show some examples in Figure 8 for the model pre-trained on 1/16 Conceptual Captions dataset and fine-tuned on 1/4 VCR dataset.

REFERENCES

Figure 1: Pseudo Ground Truth of Grounding Alignment.

Figure 2: Detailed architecture of Grounding Alignment.
Table I: **Hyper-parameter Tuning.** We do hyper-parameter tuning using $1/8$ amount of the Conceptual Caption Dataset. SPE denotes spatial positional encoding and SS Ground represents semi-supervised grounding alignment. Best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Settings</th>
<th>SPE</th>
<th>$\lambda_{\text{align}}$</th>
<th>$\lambda_{\text{gnd}}$</th>
<th>token</th>
<th>phrase</th>
<th>Visual Grounding</th>
<th>VQA</th>
<th>VCR (Q→A)</th>
<th>VCR (QA→R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLBERT [1]</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>70.92</td>
<td>67.85</td>
<td>70.83</td>
<td>72.47</td>
</tr>
<tr>
<td>+ SS Ground (token)</td>
<td>0</td>
<td>1</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>67.38</td>
<td>71.28</td>
<td>72.73</td>
</tr>
<tr>
<td>+ SS Ground (token)</td>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>67.79</td>
<td>71.77</td>
<td>73.12</td>
<td></td>
</tr>
<tr>
<td>+ SS Ground (token)</td>
<td>1</td>
<td>20</td>
<td>✓</td>
<td>70.98</td>
<td>68.45</td>
<td>71.91</td>
<td>73.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ SS Ground (phrase)</td>
<td>1</td>
<td>20</td>
<td>✓</td>
<td>67.39</td>
<td>71.52</td>
<td>72.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ SPE</td>
<td>✓</td>
<td>1</td>
<td>0</td>
<td>71.19</td>
<td>68.12</td>
<td>71.67</td>
<td>73.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ SS Ground (token)</td>
<td>✓</td>
<td>1</td>
<td>20</td>
<td>69.02</td>
<td>67.64</td>
<td>71.72</td>
<td>73.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ SS Ground (phrase)</td>
<td>✓</td>
<td>1</td>
<td>20</td>
<td>72.23</td>
<td>68.98</td>
<td>71.88</td>
<td>73.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table II: **Adjustment of Learning Rates.** We conduct the learning adjustment on the full dataset and fine-tune to visual grounding and VQA tasks. The setting is using our final model which includes SPE and semi-supervised grounding at the phrase level. Best results for each task are bold.

<table>
<thead>
<tr>
<th>Tasks / Learning Rate</th>
<th>lr=2.5e-05</th>
<th>lr=5e-05</th>
<th>lr=1e-04 (original)</th>
<th>lr=2e-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual Grounding</td>
<td>71.74</td>
<td>71.83</td>
<td><strong>72.47</strong></td>
<td>71.45</td>
</tr>
<tr>
<td>VQA</td>
<td>68.82</td>
<td>69.19</td>
<td><strong>69.63</strong></td>
<td>68.02</td>
</tr>
</tbody>
</table>
Figure 3: Qualitative Results for visual grounding task pre-trained on full Conceptual Captions dataset and fine-tuned on full RefCOCO+ dataset. Orange box: Ground truth. Red box: Baseline (ViLBERT [1]). Blue box: Ours. The top two rows are the positive examples and the bottom row are the failure examples.
Figure 4: Qualitative Results for visual grounding task pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 RefCOCO+ dataset. Orange box: Ground truth. Red box: Baseline (ViLBERT [1]). Blue box: Ours. The top two rows are the positive examples and the bottom row are the failure examples.
Figure 5: Qualitative Results for VQA task pre-trained on full Conceptual Captions dataset and fine-tuned on full VQA dataset. The top row is the positive examples and the two on the right of the bottom row are failure cases.

Figure 6: Qualitative Results for VQA task pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 VQA dataset. The top row is the positive examples and the two on the right of the bottom row are failure cases.
Figure 7: Qualitative Results for VCR task pre-trained on full Conceptual Captions dataset and fine-tuned on full VCR dataset. We show the ground truth answers with green color on the multiple choices side. The green/red colors on the probability side mean the correct/incorrect answers. The top row is the positive examples and the one on the right of the bottom row is the failure case.
Figure 8: Qualitative Results for VCR task pre-trained on 1/16 Conceptual Captions dataset and fine-tuned on 1/4 VCR dataset. We show the ground truth answers with green color on the multiple choices side. The green/red colors on the probability side mean the correct/incorrect answers. The top row is the positive examples and the one on the right of the bottom row is the failure case.