CONDITIONED AND COMPOSED IMAGE RETRIEVAL COMBINING AND PARTIALLY FINE-TUNING CLIP-BASED FEATURES

Workshop on Open-Domain Retrieval Under Multi-Modal Settings, CVPR 2022, New Orleans

1,2 Alberto BALDRATI, 1 Marco BERTINI, 1 Tiberio URICCHIO, 1 Alberto DEL BIMBO
[name.surname@unifi.it]

1 Università degli Studi di Firenze - MICC
2 Università di Pisa

Firenze, Italy - Pisa, Italy
Conditioned and composed image retrieval extends traditional CBIR systems to improve their effectiveness by adding user feedback.

I want a similar one but blue with a different character.

Add two more puppies and change the breed.
To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:
To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

1. We fine-tune the CLIP text encoder using a simple combination of visual and textual features
To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

1. We fine-tune the CLIP text encoder using a simple combination of visual and textual features

2. We train from scratch a Combiner network that learns to fuse the partially fine-tuned multimodal features
To address the conditioned and composed image retrieval tasks we propose a two-stage approach based on CLIP [1] multimodal features:

1. We fine-tune the CLIP text encoder using a simple combination of visual and textual features

2. We train from scratch a Combiner network that learns to fuse the partially fine-tuned multimodal features

**First Stage**

**Text encoder fine-tuning**

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large-scale image-text pre-training and the downstream task.
**First Stage**

**Text encoder fine-tuning**

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large-scale image-text pre-training and the downstream task.
First Stage
Text encoder fine-tuning

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task.
In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task.
First Stage
Text encoder fine-tuning

In this stage we perform a fine-tuning of the CLIP text encoder to reduce the task mismatch between the large scale image-text pre-training and the downstream task.
In this second stage we train from scratch a Combiner network that learns to combine the multimodal query features.
In this second stage we train from scratch a Combiner network that learns to combine the multimodal query features.
The Combiner network outputs a normalized sum of multiple components: a convex combination of text and image features and a learned text-image mixture.
### Comparison with sOTa

**FashionIQ dataset**

<table>
<thead>
<tr>
<th>Method</th>
<th>Shirt (R@10)</th>
<th>Shirt (R@50)</th>
<th>Dress (R@10)</th>
<th>Dress (R@50)</th>
<th>Toptee (R@10)</th>
<th>Toptee (R@50)</th>
<th>Average (R@10)</th>
<th>Average (R@50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARTEMIS [4]</td>
<td>21.78</td>
<td>43.64</td>
<td>27.16</td>
<td>52.40</td>
<td>29.20</td>
<td>54.83</td>
<td>26.05</td>
<td>50.29</td>
</tr>
<tr>
<td>RTIC-GCN w/GloVe [5]</td>
<td>23.79</td>
<td>47.25</td>
<td>29.15</td>
<td>54.04</td>
<td>31.61</td>
<td>57.98</td>
<td>28.18</td>
<td>53.09</td>
</tr>
<tr>
<td>CoSMo [6]</td>
<td>24.90</td>
<td>49.18</td>
<td>25.64</td>
<td>50.30</td>
<td>29.21</td>
<td>57.46</td>
<td>26.58</td>
<td>52.31</td>
</tr>
<tr>
<td>DCNet [8]</td>
<td>23.95</td>
<td>47.30</td>
<td>28.95</td>
<td>56.07</td>
<td>30.44</td>
<td>58.29</td>
<td>27.78</td>
<td>53.89</td>
</tr>
<tr>
<td>SAC w/BERT [9]</td>
<td>28.02</td>
<td>51.86</td>
<td>26.52</td>
<td>51.01</td>
<td>32.70</td>
<td>61.23</td>
<td>29.08</td>
<td>54.70</td>
</tr>
<tr>
<td>Baldrati et al (RN50x4)[10]</td>
<td>35.76</td>
<td>56.20</td>
<td>27.20</td>
<td>53.57</td>
<td>36.31</td>
<td>61.14</td>
<td>33.09</td>
<td>56.99</td>
</tr>
<tr>
<td>Proposed approach (RN50)</td>
<td>35.77</td>
<td>57.02</td>
<td>31.73</td>
<td>56.02</td>
<td>36.46</td>
<td>62.77</td>
<td>34.65</td>
<td>58.60</td>
</tr>
<tr>
<td>Proposed approach (RN50x4)</td>
<td><strong>39.99</strong></td>
<td><strong>60.45</strong></td>
<td><strong>33.81</strong></td>
<td><strong>59.40</strong></td>
<td><strong>41.41</strong></td>
<td><strong>65.37</strong></td>
<td><strong>38.32</strong></td>
<td><strong>61.74</strong></td>
</tr>
</tbody>
</table>

**Table:** Comparison between our method and current state-of-the-art models on the Fashion-IQ dataset. Best scores are highlighted in bold, second-best scores are underlined.
### Comparison with sOTA

**CIRR dataset**

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall@K</th>
<th>R_subset_@K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K = 1</td>
<td>K = 5</td>
</tr>
<tr>
<td>TIRG(^\dagger) [11]</td>
<td>14.61</td>
<td>48.37</td>
</tr>
<tr>
<td>MAAF(^\dagger) [12]</td>
<td>10.31</td>
<td>33.03</td>
</tr>
<tr>
<td>MAAF+BERT(^\dagger) [12]</td>
<td>10.12</td>
<td>33.10</td>
</tr>
<tr>
<td>ARTEMIS [4]</td>
<td>16.96</td>
<td>46.10</td>
</tr>
<tr>
<td>CIRPLANT(^\dagger) [3]</td>
<td>15.18</td>
<td>43.36</td>
</tr>
<tr>
<td>CIRPLANT w/OSCAR(^\dagger) [3]</td>
<td>19.55</td>
<td>52.55</td>
</tr>
<tr>
<td>Proposed approach (RN50)</td>
<td>35.81</td>
<td>68.80</td>
</tr>
<tr>
<td>Proposed approach (RN50x4)</td>
<td>38.53</td>
<td>69.98</td>
</tr>
</tbody>
</table>

**Table:** Comparison between our method and current state-of-the-art models on the CIRR test set. Best scores are highlighted in bold, second-best scores are underlined. \(^\dagger\) denotes results cited from [3]
Scan the QR Code to try a LIVE DEMO
References


