Problem Scenario

- Composed image retrieval (CIR) aims to train a model that can fuse multi-modal information to accurately retrieve images that match the query.

The Dataset Construction Procedure

- **Template-based**
  - Reference Image Caption: A dog in the water
  - Predetermined Templates and Rules
  - Relative Caption: Change a dog with a cat.
    - Target Image Caption: A cat in the water.
  - Encoders to extract features from visual and textual inputs respectively;
  - A Transformer module to capture the interaction between two modalities;
  - An adaptive aggregation module that combats modal redundancy and fuses the features together.

- **Large Language Model-based**
  - Reference Image Caption: A panda is lying on a trunk
  - Relative Caption: Change an animal to woman
    - Target Image Caption: Two pandas are lying on a trunk
  - Datasets: we initiate a scalable pipeline to automatically construct datasets for training CIR model, by simply exploiting a large-scale dataset of image-text pairs.
  - Architecture: we introduce TransAgg, which employs a simple yet efficient fusion mechanism, to adaptively combine information from diverse modalities;
  - Results: our proposed approach either performs on par with or significantly outperforms the existing state-of-the-art (SOTA) models.

Our Contribution

- **Datasets**: we initiate a scalable pipeline to automatically construct datasets for training CIR model, by simply exploiting a large-scale dataset of image-text pairs.
- **Architecture**: we introduce TransAgg, which employs a simple yet efficient fusion mechanism, to adaptively combine information from diverse modalities;
- **Results**: our proposed approach either performs on par with or significantly outperforms the existing state-of-the-art (SOTA) models.

TransAgg (Architecture)

- Datasets: we initiate a scalable pipeline to automatically construct datasets for training CIR model, by simply exploiting a large-scale dataset of image-text pairs.
- Architecture: we introduce TransAgg, which employs a simple yet efficient fusion mechanism, to adaptively combine information from diverse modalities;
- Results: our proposed approach either performs on par with or significantly outperforms the existing state-of-the-art (SOTA) models.

Experiment Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Zero-shot</th>
<th>Triples</th>
<th>R@10</th>
<th>R@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-shot</td>
<td>25.83</td>
<td>51.78</td>
<td>73.80</td>
<td>84.20</td>
</tr>
<tr>
<td>Our Approach</td>
<td>24.70</td>
<td>47.70</td>
<td>64.70</td>
<td>74.30</td>
</tr>
<tr>
<td>Sentence Transformer</td>
<td>19.76</td>
<td>37.25</td>
<td>52.81</td>
<td>62.51</td>
</tr>
</tbody>
</table>

- On CIRR dataset, our proposed model achieves state-of-the-art results in all metrics except for Recall@50.
- On FashionIQ dataset, our proposed TransAgg model trained on the automatically constructed dataset also falls among the top 2 best models.

Explainability Heatmaps

- Problem Scenario: Composed image retrieval (CIR) aims to train a model that can fuse multi-modal information to accurately retrieve images that match the query.
- Our Contribution: Datasets: we initiate a scalable pipeline to automatically construct datasets for training CIR model, by simply exploiting a large-scale dataset of image-text pairs.
- Architecture: we introduce TransAgg, which employs a simple yet efficient fusion mechanism, to adaptively combine information from diverse modalities;
- Results: our proposed approach either performs on par with or significantly outperforms the existing state-of-the-art (SOTA) models.