Understanding Art through Multi-Modal Retrieval in Paintings

Motivation
We aim to understand art beyond its visual appearance.

Can we identify the painting an artistic comment belongs to?
• We capture semantics with robust representations.
• Contextual information informs the model to boost results.

SemArt Dataset
➢ Corpus with 21,384 reproductions of European paintings.
➢ Painting associated to natural language comment.
➢ With attributes: author, title, date, techniques, type, school, timeframe.
➢ Evaluation as multimodal retrieval task:
  text → image and image → text

Multi-Modal Representations in Art
Language representation:
1. Comments and titles encoded as TFIDF vectors.
2. Language attributes encoded as one-hot vector.

Visual representation:
1. Paintings encoded with a ResNet50.
2. Visual attributes predicted with ContextNet.

Evaluation

<table>
<thead>
<tr>
<th></th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>MR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vis&amp;Lang</td>
<td>0.164</td>
<td>0.384</td>
<td>0.505</td>
<td>10</td>
<td>0.162</td>
<td>0.366</td>
<td>0.479</td>
<td>12</td>
</tr>
<tr>
<td>Att TYPE</td>
<td>0.178</td>
<td>0.383</td>
<td>0.525</td>
<td>9</td>
<td>0.165</td>
<td>0.364</td>
<td>0.491</td>
<td>11</td>
</tr>
<tr>
<td>Att SCHOOL</td>
<td>0.192</td>
<td>0.386</td>
<td>0.507</td>
<td>10</td>
<td>0.163</td>
<td>0.364</td>
<td>0.484</td>
<td>12</td>
</tr>
<tr>
<td>Att TF</td>
<td>0.127</td>
<td>0.322</td>
<td>0.432</td>
<td>18</td>
<td>0.130</td>
<td>0.336</td>
<td>0.444</td>
<td>16</td>
</tr>
<tr>
<td>Att AUTHOR</td>
<td>0.236</td>
<td>0.451</td>
<td>0.572</td>
<td>7</td>
<td>0.204</td>
<td>0.440</td>
<td>0.535</td>
<td>8</td>
</tr>
</tbody>
</table>

Knowledge Graph (KG) with paintings and attributes.
Paintings in KG encoded into graph embeddings (GE).
GE enhance ResNet features during training.
Enhanced ResNet used to predict artistic attributes.

References