Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions

Prajwal Gatti\textsuperscript{1}, Kshitij Gopal Parikh\textsuperscript{1}, Dhriti Prasanna Paul\textsuperscript{1}
Manish Gupta\textsuperscript{2}, Anand Mishra\textsuperscript{1}

\{pgatti, parikh.2, paul.4, mishra\}@iitj.ac.in, gmanish@microsoft.com
\textsuperscript{1}Indian Institute of Technology, Jodhpur; \textsuperscript{2}Microsoft
Problem of Complex Queries

• How to find this? chipmunk, badger, weasel, mongoose, or skunk?

• Complex queries
  • “difficult-to-name but easy-to-draw” objects.
  • “difficult-to-sketch but easy-to-verbalize” object’s attributes or interaction with the scene.
  • Query: “numbat digging in the ground”
Related Work

• Sketch-Based Image Retrieval (SBIR)
  • Methods: CNNs, Transformer-based methods, Deep Siamese models with triplet loss
  • Specialized forms: Zero Shot-SBIR, Finegrained SBIR, Category-level SBIR

• Text-Based Image Retrieval (TBIR)
  • Alignment of (query text, images) using VisualBERT, ViLT
  • Cross-attention-based models
  • Object tags in images
  • Contrastive learning methods, zero-shot learning methods

• Multimodal Query Based Image Retrieval
  • Reference images and category text for image retrieval.
  • speech and mouse traces as the query
  • Detailed sketch and text input
    • e-commerce product images using CNNs and LSTMs
    • scene images using CLIP

<table>
<thead>
<tr>
<th>Query</th>
<th>Dataset</th>
<th>Sketch</th>
<th>Text</th>
<th>Target Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketch</td>
<td>TU-Berlin</td>
<td>Object</td>
<td>None</td>
<td>Focused Object</td>
</tr>
<tr>
<td>Sketch</td>
<td>QMUL-Shoe-V2</td>
<td>Object</td>
<td>None</td>
<td>Focused Object</td>
</tr>
<tr>
<td>Text</td>
<td>COCO</td>
<td>None</td>
<td>Complete</td>
<td>Complete Scene</td>
</tr>
<tr>
<td>Text</td>
<td>Flickr-30K</td>
<td>None</td>
<td>Complete</td>
<td>Complete Scene</td>
</tr>
<tr>
<td>Sketch+Text</td>
<td>FS COCO</td>
<td>Scene</td>
<td>Complete</td>
<td>Complete Scene</td>
</tr>
<tr>
<td>Sketch+Text</td>
<td>CSTBIR (Ours)</td>
<td>Object</td>
<td>Complementary</td>
<td>Complete Scene</td>
</tr>
</tbody>
</table>

Given: a hand-drawn sketch $S$, a complementary text $T$ and a database $D$ of $N$ natural scene images with multiple objects.

Rank: $N$ images according to relevance to composite $\langle S, T \rangle$ query.

Natural images and text descriptions from Visual Genome.

Sketches from Quick, Draw!

Train (~1.89M queries, ~97K images)

Validation (~5K images, ~97K queries)

Test
- Test-1K: 1K queries, 1K images
- Test-5K: 4K queries, 5K images
- Open-Category set: 750 queries, 70 objects, 1K images.

Property | Value
--- | ---
Average sentence length (in words/tokens) | 5.4 / 7.7
Number of Unique Images | 108K
Number of Unique Sketches | 562K
Number of Unique Object Categories | 258
Number of Training Instances | 1.89M
Number of Validation Instances | 97K
Number of Test Instances | 5000
Avg % Area Covered by Query | 36.7

markhor, bodhran, and penny-farthing
Sketches

markhor  bouzouki  marimba
flame lily  sugarglider  jerboa
platypus  froe  echidna
pawpaw  skycar  balalaika
STNet Model for CSTBIR

• Query (Sketch+Text) Encoding
  • pretrained CLIP text encoder
  • pretrained Vision Transformer (ViT) finetuned on sketches.

• Image Encoding
  • pretrained CLIP-ViT

• STNet Training
  • $L_{CT}$: Contrastive Training (query, image)
    • InfoCNE($h^T_{CLS}, h^I_{AVG}$)
  • $L^T_{CLS}$ and $L^I_{CLS}$: Object Classification
  • $L_{OD}$: Sketch-Guided Object Detection
    • intersection over union (IoU)
  • $L_{SR}$: Sketch Reconstruction
    • eight blocks of Convolution-BatchNormReLU
    • Binary Cross Entropy loss and the DICE loss

\[
\alpha_{IS} = \text{Softmax}(\tilde{H}^I \times h^S_{CLS})
\]
\[
H^I = \alpha_{IS} \odot \tilde{H}^I
\]
\[
h^I_{AVG} = \frac{1}{m} \sum_{i=1}^{m} H^I_i
\]
Image retrieval results on Test-1K (T1K) and Test-5K (T5K)

- **Sketch-based Image Retrieval (SBIR)**
  - Doodle2Search and DeepSBIR
  - ViT-based Siamese: 2 ImageNet pre-trained ViT encoders for sketch and image modalities trained using InfoNCE

- **Text-based Image Retrieval (TBIR)**
  - VisualBERT, ViLT, CLIP

- **Composite Query-based Image Retrieval**
  - TIRG and Taskformer
  - 2-stage
    - ViT trained for sketch classification to get an object name
    - Insert(object name, incomplete text query) and use pretrained CLIP
  - 2-stage (desc): Insert(object description, incomplete text query)

<table>
<thead>
<tr>
<th>Method</th>
<th>R@10↑</th>
<th>R@20↑</th>
<th>R@50↑</th>
<th>R@100↑</th>
<th>MdR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1K</td>
<td>T5K</td>
<td>T1K</td>
<td>T5K</td>
<td>T1K</td>
</tr>
<tr>
<td>Sketch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doodle2Search</td>
<td>14.3</td>
<td>3.6</td>
<td>24.5</td>
<td>6.7</td>
<td>36.2</td>
</tr>
<tr>
<td>DeepSBIR</td>
<td>5.2</td>
<td>1.6</td>
<td>8.8</td>
<td>3.0</td>
<td>18.9</td>
</tr>
<tr>
<td>ViT-Siamese</td>
<td>20.4</td>
<td>5.2</td>
<td>34.2</td>
<td>9.9</td>
<td>51.0</td>
</tr>
<tr>
<td>Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VisualBERT</td>
<td>23.3</td>
<td>7.6</td>
<td>35.9</td>
<td>15.4</td>
<td>40.8</td>
</tr>
<tr>
<td>ViLT</td>
<td>28.1</td>
<td>10.5</td>
<td>42.7</td>
<td>16.5</td>
<td>60.2</td>
</tr>
<tr>
<td>CLIP</td>
<td>50.6</td>
<td>24.2</td>
<td>63.1</td>
<td>33.7</td>
<td>78.8</td>
</tr>
<tr>
<td>Sketch+Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIRG</td>
<td>31.9</td>
<td>10.4</td>
<td>44.2</td>
<td>17.3</td>
<td>62.8</td>
</tr>
<tr>
<td>Taskformer</td>
<td>22.4</td>
<td>9.3</td>
<td>35.6</td>
<td>14.8</td>
<td>42.3</td>
</tr>
<tr>
<td>Two-stage</td>
<td>67.0</td>
<td>34.8</td>
<td>77.7</td>
<td>46.9</td>
<td>88.6</td>
</tr>
<tr>
<td>Two-stage (desc)</td>
<td>60.1</td>
<td>30.5</td>
<td>73.7</td>
<td>41.7</td>
<td>85.5</td>
</tr>
<tr>
<td>STNET (Ours)</td>
<td>73.7</td>
<td>38.7</td>
<td>80.6</td>
<td>50.0</td>
<td>89.4</td>
</tr>
</tbody>
</table>

- **sketch+text > text-only > sketch-only**
- **STNet>2-stage**
  - incomplete semantics in object name
  - Ambiguous objects: mouse, bat, star

Further Experiments and Results

- **Ablation study on Test-1K**
  - Removing any of the 3 losses hurts.
  - Removing $L_{CLS}$ hurts the most.

- **Results on Open-Category Test Set**
  - Open-Category setting is difficult.
  - STNet is more robust to this complex setting.

<table>
<thead>
<tr>
<th>M</th>
<th>Query</th>
<th>Objective</th>
<th>R@10</th>
<th>R@20</th>
<th>R@50</th>
<th>R@100</th>
<th>MdR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>$\mathcal{L}_{CT}$</td>
<td>20.2</td>
<td>33.7</td>
<td>50.9</td>
<td>62.9</td>
<td>50.5</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>$\mathcal{L}_{CT}$</td>
<td>50.6</td>
<td>63.1</td>
<td>78.8</td>
<td>86.7</td>
<td>10.0</td>
</tr>
<tr>
<td>3</td>
<td>T+S</td>
<td>$\mathcal{L}_{CT}$</td>
<td>68.4</td>
<td>77.2</td>
<td>85.6</td>
<td>89.8</td>
<td>5.0</td>
</tr>
<tr>
<td>4</td>
<td>T+S</td>
<td>$\mathcal{L}<em>{CT} + \mathcal{L}</em>{OD} + \mathcal{L}_{SR}$</td>
<td>69.4</td>
<td>80.4</td>
<td>85.6</td>
<td>90.4</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>T+S</td>
<td>$\mathcal{L}<em>{CT} + \mathcal{L}</em>{CLS} + \mathcal{L}_{SR}$</td>
<td>70.4</td>
<td>79.6</td>
<td>86.2</td>
<td>91.1</td>
<td>5.0</td>
</tr>
<tr>
<td>6</td>
<td>T+S</td>
<td>$\mathcal{L}<em>{CT} + \mathcal{L}</em>{CLS} + \mathcal{L}_{OD}$</td>
<td>71.2</td>
<td>79.0</td>
<td>87.0</td>
<td>93.0</td>
<td>4.0</td>
</tr>
<tr>
<td>7</td>
<td>T+S</td>
<td>$\mathcal{L}<em>{CT} + \mathcal{L}</em>{CLS} + \mathcal{L}<em>{OD}$ + $\mathcal{L}</em>{SR}$</td>
<td><strong>73.7</strong></td>
<td><strong>80.6</strong></td>
<td><strong>89.4</strong></td>
<td><strong>93.5</strong></td>
<td><strong>3.0</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>R@10↑</th>
<th>R@25↑</th>
<th>R@50↑</th>
<th>R@100↑</th>
<th>MdR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-Siamese</td>
<td>6.3</td>
<td>8.6</td>
<td>14.5</td>
<td>23</td>
<td>241.0</td>
</tr>
<tr>
<td>CLIP</td>
<td>21.6</td>
<td>30.6</td>
<td>39.4</td>
<td>47.6</td>
<td>71.0</td>
</tr>
<tr>
<td>Two-Stage</td>
<td>29.0</td>
<td>38.2</td>
<td>48.8</td>
<td>54.8</td>
<td>63.0</td>
</tr>
<tr>
<td>STNET (Ours)</td>
<td><strong>37.2</strong></td>
<td><strong>45.3</strong></td>
<td><strong>62.3</strong></td>
<td><strong>71.7</strong></td>
<td><strong>27.5</strong></td>
</tr>
</tbody>
</table>
capybara, sitar, penny-farthing, and okapi.

Search Queries

- Capybara on a slide being fed red ice cream
- Bearded man on the bank of a river playing sitar besides a man playing tabla.
- Person dressed in a suit standing beside a penny-farthing.
- Pair of okapis feeding on green grass.

Top-5 Retrieved Results

Conclusion

• CSTBIR (Composite Sketch+Text Based Image Retrieval)
  • New dataset: ~2M queries and ~108K natural scene images.
  • STNet (Sketch+Text Network)
    • Pretrained multimodal transformer
    • Uses a hand-drawn sketch to localize relevant objects in the natural scene image
    • Encodes the text and image to perform image retrieval
    • contrastive loss, object classification loss, sketch-guided object detection loss, and sketch reconstruction loss

• Search for missing people, search for a product in digital catalogs, …

• Thanks!
  • LinkedIn: http://aka.ms/manishgupta
  • HomePage: https://sites.google.com/view/manishg/