Semantics to Space (S2S): Embedding semantics into spatial space for zero-shot verb-object query inferencing

Sungmin Eum*†
eum_sungmin@bah.com

Heesung Kwon*
heesung.kwon.civ@mail.mil

*U.S. Army Research Laboratory
†Booz Allen Hamilton
Semantics to Space(S2S): Embedding semantics into spatial space for **zero-shot** verb-object query inferencing

**Verb-object query inferencing?**
Recognize images by using verb-object expressions, e.g., “ride a horse”, “hold ball”

**Zero-shot?**
Be able to classify/recognize images that fall under “unseen” category
**TASK:** Zero-shot Verb-Object (VO) Inferencing

Seen VO (TRAIN)

- hold + bat
- hold + bottle
- hold + cup

Unseen VO (TEST)

- hold + ball
- hold + orange
- ride + bicycle
- ride + cow
- ride + elephant
- ride + horse
- ride + motorcycle

Generalize “hold” & “ride”

* [ ball, orange, horse, motorcycle ] was never shown in training.
Comparison with previous work:

**Previous work**

![Diagram of previous work with four types: Type I, Type II, Type III, Type IV.]

**Ours**

- Do NOT compete with previous approaches in searching for the best joint embedding subspace.

**Q1** “Why are the *semantics* only exploited for constructing query features but *not* for generating visual features when the end goal is to train a module to match the two?”

**Q2** “Will the semantics be effective if embedded into both the visual and the semantic stream?”

- Directly embed the semantics into the spatial space of the visual stream (Contribution).

**Competition:** What is the best joint embedding subspace for co-learning visual features with query/semantic features?
**APPROACH:** Semantics-to-Space (S2S)
**DATASET: Verb-Transferability 60**

<table>
<thead>
<tr>
<th>Verbs</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>eat</td>
<td>apple, banana</td>
<td>broccoli, donut</td>
</tr>
<tr>
<td>feed</td>
<td>bird, cat, cow</td>
<td>dog, giraffe, horse, sheep</td>
</tr>
<tr>
<td>hold</td>
<td>baseball bat, book, bottle,</td>
<td>orange, scissors, skateboard,</td>
</tr>
<tr>
<td></td>
<td>carrot, cell phone, cup, frisbee, hair dryer, handbag, knife</td>
<td>sports ball, surfboard, tennis racket, toothbrush, vase, wine glass</td>
</tr>
<tr>
<td>kiss</td>
<td>dog, giraffe, horse</td>
<td>toothbrush, vase, wine glass</td>
</tr>
<tr>
<td>lie on</td>
<td>bed, bench</td>
<td>bird, cat, cow</td>
</tr>
<tr>
<td>ride</td>
<td>bicycle, cow, elephant</td>
<td>couch, surfboard</td>
</tr>
<tr>
<td>sit on</td>
<td>bed, bench</td>
<td>horse, motorcycle, sheep</td>
</tr>
<tr>
<td>stand on</td>
<td>bed, bench</td>
<td>chair, couch</td>
</tr>
<tr>
<td>wash</td>
<td>bicycle, cow, dog</td>
<td>elephant, horse, motorcycle</td>
</tr>
</tbody>
</table>
EXPERIMENT: Scenario 1. Verb-Transferability

Evaluate how well the network transfers the “seen” verbs paired with “unseen” objects

**Seen (Train)**

- eat + apple
- eat + banana

**Unseen (Test)**

- eat + broccoli?
- feed + broccoli?
- hold + broccoli?
- kiss + broccoli?
- lie on + broccoli?
- ride + broccoli?
- sit on + broccoli?
- stand on + broccoli?
- wash + broccoli?
**EXPERIMENT:** Scenario 1. Verb-Transferability

Evaluate how well the network transfers the “seen” verbs paired with “unseen” objects

---

**Verb transferability evaluation.** BASELINE: Sung et al. [8], OrthoVec2S: Orthonormal Vectors-to-space, S2S: Semantics-to-space. All the numbers indicate recognition accuracy in [%].

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Sung et al.</th>
<th>OrthoVec2S</th>
<th>S2S</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>33.53</td>
<td>40.40</td>
<td>46.27</td>
</tr>
<tr>
<td>ResNet34</td>
<td>38.00</td>
<td>43.53</td>
<td>48.87</td>
</tr>
<tr>
<td>ResNet50</td>
<td>41.73</td>
<td>44.60</td>
<td>50.47</td>
</tr>
</tbody>
</table>
CONCLUSION:

Semantics to Space (S2S): Embedding semantics into spatial space for zero-shot verb-object query inferencing

- Introduced a simple, yet powerful semantics embedding approach for two-stream ZSL approach
- Augmented visual information by directly embedding the semantics in a spatial sense
- Validated that S2S can be used as a general module to enhance the performances of various ZSL baseline architectures

Please check out our manuscript!!