SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition

Kaiyu Yang, Olga Russakovsky, Jia Deng
Department of Computer Science, Princeton University

1. Introduction

Dataset bias: easy spatial relations which can be guessed without looking at the image

- onion ON table
- peach ON table
- apple ON table
- jar ON table
- tomatoes ON table
- cup ON table

SpatialSense: a visual relationship dataset with more challenging examples

- bull in front of kid ✓
- truck on chair ✗
- glasses on man ✗
- soldier above forest ✓

Constructed via adversarial crowdsourcing to reduce bias and obtain challenging examples in the long tail

2. Adversarial Crowdsourcing

Annotators identify challenging relations to confuse a “robot”

1. Given an image, propose a positive/negative spatial relation with two objects and a predicate, e.g., store on fridge.
2. The robot tries to guess whether the relation is positive or negative. You win if the robot is wrong. Otherwise, it provides feedback and you try again.

robot = language model + 2D model

3. The errors made by state-of-the-art models have high correlation with simple baselines

4. Reduced Dataset Bias

SpatialSense contains less bias than VRD [1] and VG [2]

Predicates harder to predict from language and 2D cues alone

Given object names and bounding boxes, without image pixels

1. SpatialSense is the hardest (has the lowest accuracy)

2. Models trained on SpatialSense exhibit better cross-dataset generalization [8]

Effect of adversarial crowdsourcing

SpatialSense is harder (has lower accuracy) than an ablation dataset collected without adversarial crowdsourcing.

4. Data Statistics and Distributions

17.5K relations on 11.6K images, 3.7K unique object classes with 2.1K of them appearing only once

More balanced predicate distributions than VRD [1] and VG [2]

More balanced distribution of 2D spatial location

Harder to determine the predicate from 2D cues alone

2D locations of subjects relative to objects

5. Baselines

State-of-the-art models struggle on SpatialSense

Task: Given the image, the names and bounding boxes of two objects, and a predicate, classify whether the relation holds

Methods

1. Simple baselines based solely on language and/or 2D cues
2. State-of-the-art models for visual relationship detection [3-7]

Results

1. SpatialSense is challenging: The best models perform around 70%, which is quite low for a binary classification task

2. Requires deeper visual reasoning: State-of-the-art models perform similarly to the simple 2D baseline, suggesting they might rely too much on 2D cues

Model | Overall | above | behind | in | in front | next to | on | to the left of | to the right of | under
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
Language-only | 60.1 | 59.4 | 62.0 | 56.4 | 55.1 | 58.8 | 63.2 | 57.4 | 61.1 | 59.3
Language + 2D | 71.5 | 61.1 | 67.9 | 59.2 | 60.2 | 64.6 | 79.0 | 68.9 | 73.7 | 67.9
VIP-FCN [1] | 67.3 | 56.1 | 68.1 | 56.0 | 62.7 | 73.5 | 72.3 | 68.7 | 75.3 | 66.4
PRL [3] | 67.5 | 59.0 | 67.1 | 58.4 | 70.5 | 71.1 | 63.7 | 69.2 | 66.2 | 66.2
PRL [1] | 66.3 | 61.5 | 65.2 | 57.4 | 66.2 | 72.0 | 69.1 | 71.9 | 59.3 | 66.4
DRNet [8] | 73.8 | 58.4 | 71.2 | 59.8 | 46.9 | 55.6 | 63.5 | 63.5 | 77.9 | 65.6
Hsu et al. [9] | 69.4 | 61.5 | 69.7 | 67.8 | 64.9 | 57.7 | 76.2 | 64.6 | 68.5 | 75.9
Hsu et al. [10] | 74.6 | 60.0 | 96.3 | 95.8 | 94.5 | 95.7 | 88.9 | 99.2 | 96.4 | 96.4

...