## SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition

PRINCETON<br>UNIVERSITY

Kaiyu Yang, Olga Russakovsky, Jia Deng

Department of Computer Science, Princeton University
回回


## 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., stove 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 +2 D model

## 4. Data Statistics and Distributions

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

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


Predicate distributions of the most frequent objects (Left), and top-50 objects (Right)
More balanced distribution of 2D spatial location
Harder to determine the predicate from 2D cues alone

$$
\text { - to the left of } \cdot \text { to the right of }
$$

2D locations of subjects relative to objects


## 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]

|  | VRD | vG | SpatialSense | average drop |
| :---: | :---: | :---: | :---: | :---: |
| VRD [1] | 66.9/59.6 | 45.2/53.2 | 30.6/39.9 | 29.0/13.1 |
| VG [2] | 50.8/38.2 | 76.0/65.3 | 36.1/33.8 | 32.6/29.3 |
| SpatialSense | 40.3/44.6 | 42.8/52.5 | 39.8/43.4 | -1.8/-5.2 |

Accuracies and cross-dataset generalization of the language model/2D model
Effect of adversarial crowdsourcing
SpatialSense is harder (has lower accuracy) than an ablation dataset collected without adversarial crowdsourcing.

[^0]
## 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 frot of | nexto |  | to the left | , |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| lage | ${ }_{68.8}^{60.1}$ | ${ }^{60.4}$ | ${ }_{6}^{620}$ | ${ }_{7}^{54.4}$ | ${ }^{63}$ | ${ }_{62.0}^{56.8}$ | ${ }_{76.0}^{63.2}$ | ${ }_{66.3}^{51.7}$ | ${ }_{54.7}^{54.1}$ | ${ }_{\substack{70.3 \\ 67.9}}^{6}$ |
| ${ }_{\text {20, }}^{\text {Langlage }}$ | 71.1 |  | ${ }_{675}^{669}$ | ${ }_{692}^{76.7}$ | 6.2 | 64.8 | ${ }^{77.9}$ | ${ }_{6}^{6,7}$ | ${ }_{74.7}^{74.7}$ | ${ }_{77.2}^{6.9}$ |
| $\mathrm{V}_{\text {ip-CNN }}$ | 67.2 675 | ${ }_{\text {cois }}^{59}$ | ${ }_{6}^{68.1}$ | ${ }_{6}^{6,0}$ | ${ }^{6.2}$ | ${ }_{65,7}^{62.3}$ | ${ }_{75.6}^{72.5}$ | ${ }_{56.7}^{69.7}$ | ${ }^{3.3}$ |  |
| PPR-FCN [ [] | 6.3 |  | 65.2 | 70.4 |  |  |  |  |  | 59.3 |
|  | 71.3 | 62.8 | 72.2 | ${ }^{69.8}$ | 66.9 | 59.9 | 79.4 | 63.5 | 6.4 | 75.9 |
| UTans | $\xrightarrow{694.6}$ | $\stackrel{6.5}{90 .}$ | ${ }_{96,3}$ | ${ }_{9} 9.8$ |  | 57.7 | 76.2 |  |  |  |

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


Failing examples


## Lu eral., "Visual relationship detection with language priors." ECCV. 2016.


Peyre e al., "Weakly-supervised learning of visual relations." ICCV. 2017 ".
Daie tal., "Detecting visual relationships with deep relational networks." "CVPR. 2017.
Zhang.
Thang al., V isual translation embedding network for visual relation detection." CVPR. 2017 .


[^0]:    |  | Language | 2D Locations |
    | :---: | :---: | :---: | | w/o adversariaa crowdsourcing | 69.2 | 71.3 |
    | :--- | :--- | :--- |
    | w/ adversarial crowdsourcing | $\mathbf{5 6 . 4}$ | $\mathbf{6 5 . 2}$ |

