SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition

Kaiyu Yang
Princeton University
kaiyuy@cs.princeton.edu

Olga Russakovsky
Princeton University
olgarus@cs.princeton.edu

Jia Deng
Princeton University
jiadeng@cs.princeton.edu

1. Mapping the Predicates from VRD and VG to SpatialSense

In section 4.1, in order to make the three datasets comparable, we map the spatial predicates in VRD and Visual Genome to their equivalents in SpatialSense. Here we describe the detailed mapping in Table A.

SpatialSense	above	behind	in	in front of	next to	on	to the left of	to the right of	under
VRD	above, over	behind, stand behind, sit behind, park behind	in, inside	in the front of	sleep next to, sit next to, stand next to, park next, walk next to, beside, walk beside, adjacent to	on, on the top of, sit on, stand on, drive on, park on, lying on, lean on, sleep on, rest on, skate on	on the left of	on the right of	under, stand under, sit under, below, beneath
VG	above, above a, above an, above the, are above, are above a, is above, on top of, over	behind, are behind, are behind a, behind a, behind an, behind the, is behind, is behind the, on back of	in, are in, are in a, are in an, are in the, flying in, hanging in, in a, in an, in the, inside, inside of, is in, is in a, is in the, laying in, sitting in, walking in	in front of, are in front of, are in front of a, in front of a, in front of an, in front of the, is in front of	next to, are next to, are next to a, are next to the, beside, is next to, is next to the, next to a, next to an, next to the, standing next to	on, are on, are on a, are on an, are on the, growing on, hanging on, is on, is on a, is on the, laying on, lying on, on a, on a a, on an, on are, on front of, on the, painted on, parked on, printed on, sitting on, sitting on top of, standing on, walking on, written on	are left of, in left, in left side of, left of, left side of, on left, on left of, on left side of, to left, to left of, to left of a	are to right of, on right, on right of, on right side, on right side of, right of, right side of, to right, to right of, to right of a	under, are under, are under a, are under an, below, beneath, is under, is under the, under a, under an, under the, underneath

Table A: We map the spatial predicates in VRD and Visual Genome to our predefined list of 9 predicates. We mannually check all predicates in VRD to figure out the mapping. For Visual Genome, since there is no closed vocabulary, we examined the top-100 most frequent predicates.

2. Model Architectures

We describe in details the architectures of the models used in our submission (Fig. A, B, C and D). We always add batch normalization [Ioffe and Szegedy, 2015] and ReLU [Nair and Hinton, 2010] non-linearity after each parametric layer except the output. Word embeddings are 300-dimensional and computed by a pretrained Word2Vec [Mikolov et al., 2013] model. All models are implemented using Pytorch [Paszke et al., 2017].

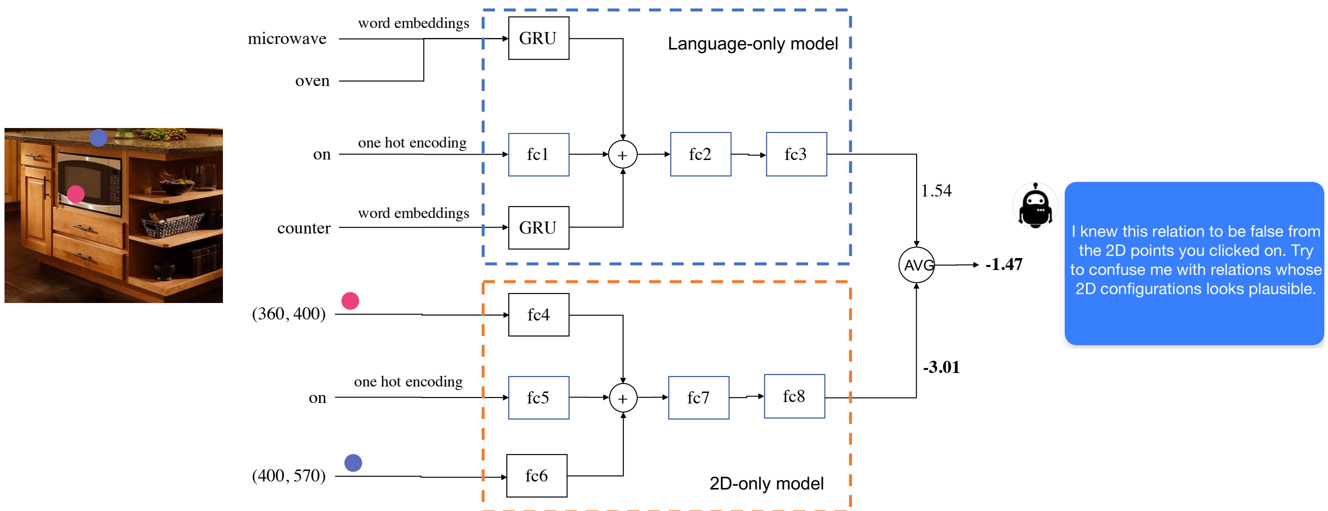
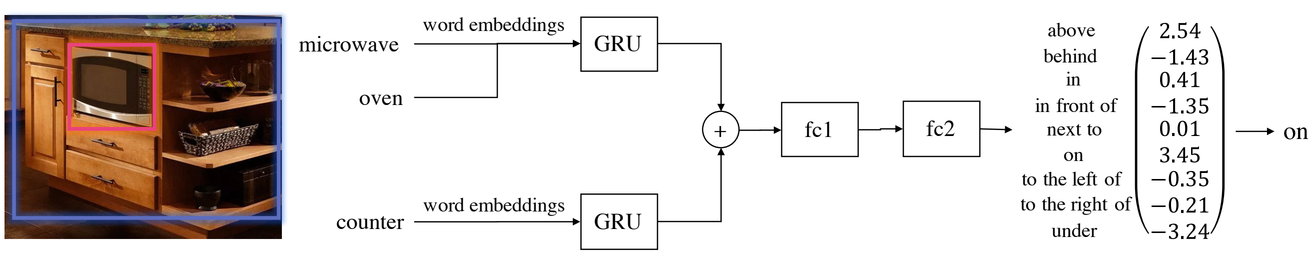
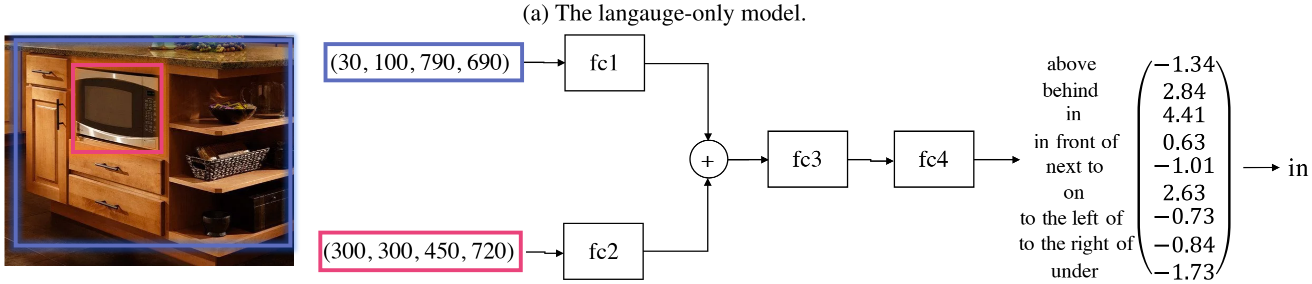


Figure A: In **adversarial crowdsourcing** (section 3 in our submission), the architecture of the robot is an ensemble of a language-only model and a 2D-only model. The language-only model takes two object names along with the predicate ("microwave oven", "on", "counter"), and outputs a score for the relation to hold (1.54). The word embeddings of object names are encoded into 512-dimensional vectors by a gated recurrent unit (GRU) [Cho et al., 2014] of 512 hidden units. The same GRU is shared between the subject and the object. The one hot encoding of the predicate is mapped to a 512-dimensional vector by a linear layer. The three feature vectors are fused by element-wise addition, on top of which a 2-layer fully connected network (with 256 hidden units) outputs the score. For the 2D-only model, linear layers map the object coordinates to 512-dimensional vectors, and others remain the same. The final output is the average of these two models.

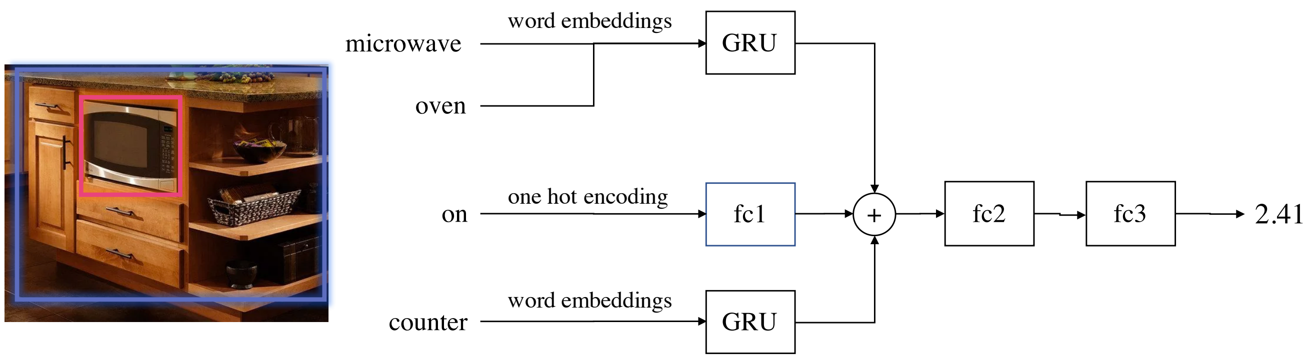


(a) The langauge-only model.

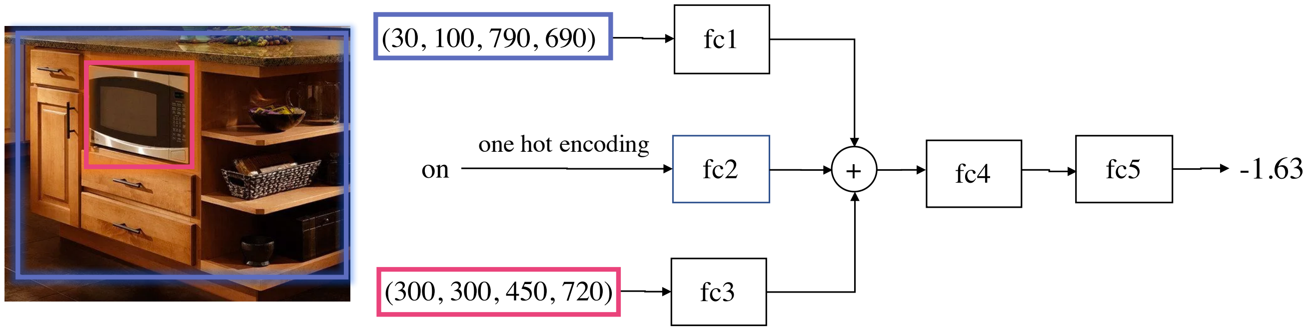


(b) The 2D-only model.

Figure B: When **classifying the predicates in *VRD-Spatial*, *VG-Spatial* and *SpatialSense-Positive*** (table 1 in our submission), we also have a language-only model and a 2D-only model. The architectures are similar to the robot; but there are three differences: (1) The branch for the input predicate is removed, since the task now is to predict the predicate. (2) The output layers now have dimension 9 instead of 1. (3) The object 2D locations are encoded by bounding boxes.

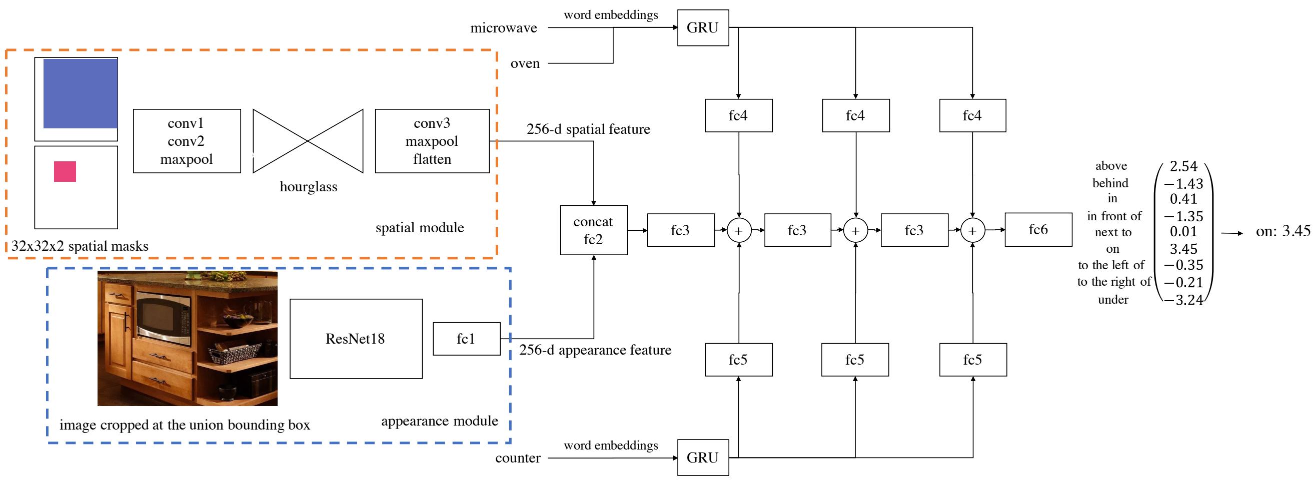


(a) The langauge-only model.

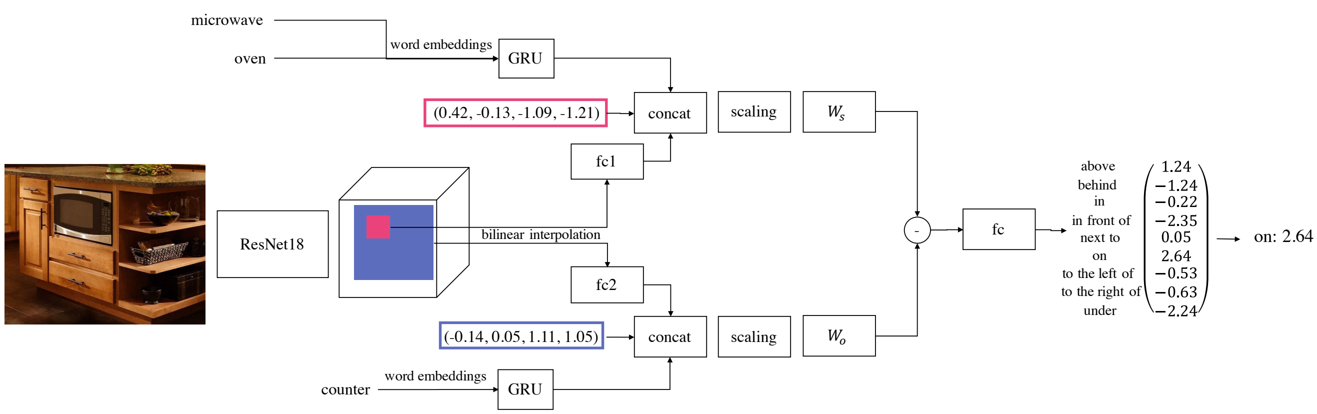


(b) The 2D-only model.

Figure C: These are the **language and 2D baselines for spatial relation recognition** (section 5 in our submission), which are also used when **quantifying the effect of adversarial crowdsourcing** (table 2 in our submission). The architectures are the same as the robot, but the object 2D locations are encoded by bounding boxes (They are annotated in a separate process and therefore not available during adversarial crowdsourcing).



(a) The **DRNet** [Dai et al., 2017] contains a spatial module and an appearance module, which respectively encode the masks of the bounding boxes and image cropped at the union bounding box into 256-dimensional feature vectors. The spatial module contains a hourglass network [Newell et al., 2016], which we find to perform better than a simple stack of convolutional layers. The appearance module is a linear layer on top of a ResNet18 [He et al., 2016] network. The spatial and appearance features as well as the object name features go through an iterative reasoning module that makes extensively use of weight-sharing; all layers with the same name (*e.g.* fc4) share the same weights. Unlike in the original DRNet paper, we do not perform iterative updates to the object name features, because they are given as ground truth in our task.



(b) For **VtransE** [Zhang et al., 2017], the bounding boxes are encoded as in the original paper. Image features are also extracted by a ResNet18 network.

Figure D: The specific instance of **DRNet and VTransE** we use for spatial relation recognition (section 5 in our submission). The input relation is *"microwave oven on counter"*. The final output is therefore the score for the predicate "on".

References

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[Dai et al., 2017] Dai, B., Zhang, Y., and Lin, D. (2017). Detecting visual relationships with deep relational networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[Newell et al., 2016] Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer.

[Paszke et al., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

[Zhang et al., 2017] Zhang, H., Kyaw, Z., Chang, S.-F., and Chua, T.-S. (2017). Visual translation embedding network for visual relation detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.