Supplementary Materials of Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation

Ruichi Yu, Ang Li, Vlad I. Morariu, Larry S. Davis
University of Maryland, College Park
{richyu,angli,morariu,lsd}@umiacs.umd.edu

1. Semantic and Spatial Representations

Spatial features have been demonstrated to be helpful in visual tasks such as object detection, image retrieval, and semantic segmentation [13, 12, 9, 6, 1, 2, 7]. For visual relationship detection task, spatial features such as the relative location and size of two objects are informative for predicate prediction. For example, relative location is a discriminative feature for predicates "under" and "above" of the VRD training set shown in Figure 2(a) and 2(b). To explicitly model the spatial location and size of an object, we use the features from [10]:

$$\left[\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}, \frac{A}{A_{img}}\right] \quad (1)$$

where W and H are the width and height of the image, A and A_{img} are the areas of the object and the image, respectively. We concatenate the above features of two objects as the spatial feature (SF) for a $\langle subj, obj \rangle$ pair.

Given a fixed $\langle subj, obj \rangle$ pair, only a few predicates are relevant, something which emerges by a simple analysis of linguistic descriptions of objects and their relationships. The strong correlation between a predicate and the $\langle subj, obj \rangle$ pairs indicates that the semantic object representation (a description of its type or class) should be informative for relationship detection, especially when the spatial features are ambiguous for different predicates. For instance, the relative bounding box locations of the subject and object w.r.t. predicate "on" and "in" as shown in 2(c) and 2(d) are similar, but given a subject of type "plate" and an object of type "table", common sense predicts "on" as the predicate.

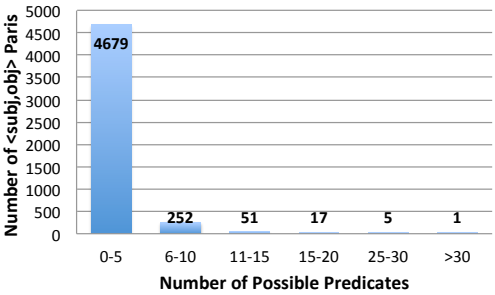
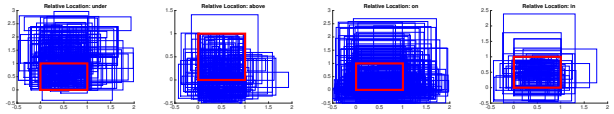


Figure 1. Number of $\langle subj, obj \rangle$ Pairs vs. Number of Possible Predicates



(a) under (b) above (c) on (d) in

Figure 2. Relative location of subjects and objects for different predicates. Red box (coordinates normalized as $[0, 0, 1, 1]$ for $[x_1, y_1, x_2, y_2]$) is the subject, blue boxes are the objects. Predicates like "under" and "above" are discriminative based on relative location. However, "on" and "in" are different to distinguish.

1.1. Evaluation of Semantic and Spatial Representations

We evaluate the use of semantic object representations and the spatial features without LK distillation. We compare our method with three methods in [8] and [11]: "Visual Phrases" denotes the method that trains deformable parts models for each relationships; "Joint CNN" denotes the method trains a 270-way CNN model to predict the subject, object and predicate together. "VRD - V only" denotes the method proposed in [8] that uses VGG-net to extract features for the BB-Union of two bounding boxes and then feeds the features into a learned linear model to predict the predicate. Besides the existing methods, we also compare our method with a baseline that uses the same VGG-net as [8] and BB-Union of boxes but with end-to-end training (denoted as "Baseline: U"). To test the use of our semantic and spatial features, we test different combinations, shown in Table 1 and Table 2.

Table 1. Predicate Detection: Different Visual Features. "U" is the union of two objects' bounding boxes; "SF" is the spatial representation; "one-hot" and "word2vec" are two semantic representations.

	R@100/50[1], k=1	R@100, k=70	R@50, k=70
Visual Phrases [11]	1.91	-	-
Joint CNN [3]	2.03	-	-
VRD - V only [8]	7.11	37.20	28.36
Baseline: U	34.82	83.15	70.02
U + one-hot	36.42	84.66	71.07
U + word2vec	37.15	83.78	70.75
U + SF	36.33	83.68	69.87
U + one-hot + SF	38.87	84.34	71.79
U + word2vec + SF	**41.33**	**84.89**	**72.29**

Table 2. Predicate Detection: Different Visual Features - Zero Shot. We use the same notations as in Table 1.

	R@100/50, k=1	R@100, k=70	R@50, k=70
VRD - V only [8]	3.52	32.34	23.95
Baseline: U	12.75	69.42	47.84
U + one-hot	11.86	69.46	47.22
U + word2vec	13.44	**69.77**	**49.01**
U + SF	**14.33**	69.01	48.32
U + one-hot + SF	12.98	69.35	48.79
U + word2vec + SF	14.13	69.41	48.13

Table 1 reveals that end-to-end training (our "Baseline: U ") with soft-max prediction outperforms the feature+linear model method ("VRD - V only"), highlighting the importance of fine-tuning. In addition, adding semantic or spatial features individually improves the predictive power of the data-driven model. The combination of BB-Union, semantic and spatial features yields the best performance.

We report the performance of different methods on zero-shot predicate prediction task in Table 2. Our end-to-end CNN model using BB-Union, semantic and spatial features outperforms [8] by a large margin in the zero-shot setting. Among the three sets of features, the spatial feature generalizes best, the one-hot feature the worst, the word2vec feature in between. We believe that this is because some $\langle subj, obj \rangle$ pairs in the zero-shot setting never occur in the training set, and the one-hot feature is too specific to generalize from the seen pairs to the unseen pairs. The word2vec features generalize better because the formulation of the vector space already takes the semantic similarity of different objects into consideration. Spatial features that capture low level information of bounding boxes generalize the best.

[1]In predicate detection task, R@100,k=1 and R@50,k=1 are exactly equivalent because there aren't enough ground truth objects in the image to produce more than 50 predictions, just as in [8].

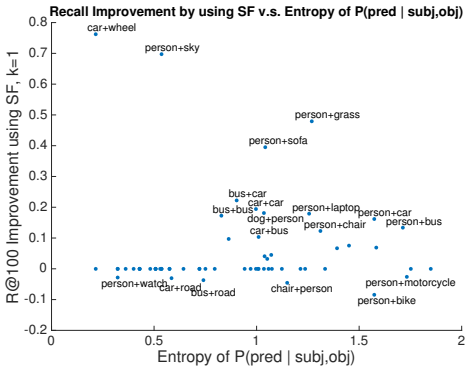


Figure 3. R@100, k=1 improvement vs. Entropy of $P(pred|subj, obj)$

1.2. When do Spatial Features Help?

Figure 1 shows that most of the $\langle subj, obj \rangle$ pairs have low entropy in their predicate distribution, which implies a strong correlation between a predicate and $\langle subj, obj \rangle$ pairs. It is interesting to understand why adding the spatial features outperforms a model that already knows which objects it is analyzing. We choose the method "U + word2vec + SF" and "U + word2vec" to evaluate when spatial features help. Our hypothesise is that SF helps when $\langle subj, obj \rangle$ is not very deterministic. We plot the "R@100/50, k=1" improvement of "U + word2vec + SF" over "U + word2vec" vs. the entropy of $P(pred|subj, obj)$ in Figure 3.

We notice that when the entropy is small ($\langle subj, obj \rangle$ is deterministic), most of the relationships with those $\langle subj, obj \rangle$ pairs do not benefit from adding spatial features. However, when the entropy is large (1-1.5), adding spatial features improves predictive power for those relationships.

2. Linguistic Knowledge Distillation with More Training Images

Another interesting finding is that LK-distillation offers higher performance improvement for the relationships that have few training instances, which is also reported in [4] (details can be found in the supplementary materials). This observation supports our motivation that knowledge helps the long-tail relationships most.

A natural question is: can we just collect more data and obtain significant improvement in predictive power and generalization? To answer this question, we utilize the recently proposed Visual Genome dataset [5] to augment the training data of VRD. By selecting the relationship instances and have predicates and objects in the categories of VRD dataset from images that are not in VRD set from Visual Genome set, we obtain around 130K more training instances, enlarging the original VRD training set by 5×. To make the zero-shot testing set of VRD remain unseen, we ensure that the relationships in the zero-shot set

are excluded from the augmented training data. We conduct similar experiments with the augmented dataset and distill the same linguistic knowledge extracted from VRD dataset. The results are shown in the Part 2 of Table 1, 2 and 3 in the main paper. We observe that training with more data leads to only marginal performance improvement of almost all baselines and proposed methods. However, for all experimental settings, with more data, our LK distillation framework still brings significant improvements, and the combination of the teacher and student networks still yields the best performance. The experiments imply that simply adding more training images does not help much. One reason we observe is that the distribution of relationships in the newly added images are still long-tail (although the Visual Genome dataset is much larger than VRD), the data-driven model still does not get enough training data. Another reason is that since we still use the linguistic knowledge extracted from VRD set, the strong linguistic prior dominates the predictions so that we observe the performance of teacher networks are very similar with/without the newly added images from Visual Genome dataset. Both reasons imply that we may need to extract the linguistic knowledge from a larger domain that contains more unseen relationships.

References

[1] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Comput. Vis. Image Underst.*, 114(6):712–722, June 2010. 1

[2] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2008. 1

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[4] Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing. Deep neural networks with massive learned knowledge. In *EMNLP, Austin, Texas, USA, November 1-4, 2016*, pages 1670–1679, 2016. 2

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 2

[6] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pages 239–253, Berlin, Heidelberg, 2010. Springer-Verlag. 1

[7] A. Li, J. Sun, J. Y. Ng, R. Yu, V. I. Morariu, and L. S. Davis. Generating holistic 3d scene abstractions for text-based image retrieval. *CoRR*, abs/1611.09392, 2016. 1

[8] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 1, 2

[9] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[10] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[11] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. 2011. 1, 2

[12] R. Salakhutdinov, A. Torralba, and J. B. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, pages 1481–1488, 2011. 1

[13] R. Yu, X. Chen, V. I. Morariu, and L. S. Davis. The role of context selection in object detection. In *British Machine Vision Conference (BMVC)*, 2016. 1