Context for Object Detection via Lightweight Global and Mid-level Representations

Mesut Erhan Unal Adriana Kovashka

University of Pittsburgh

Introduction

- Context is an important mechanism that makes visual recognition easy for humans [Palmer, 1975, Biederman et al., 1982] hence it is natural to also model context in machine perception.
- State-of-the-art two-stage object detection frameworks (e.g. Faster R-CNN) classify each region in isolation, without considering what is in the rest of the image.

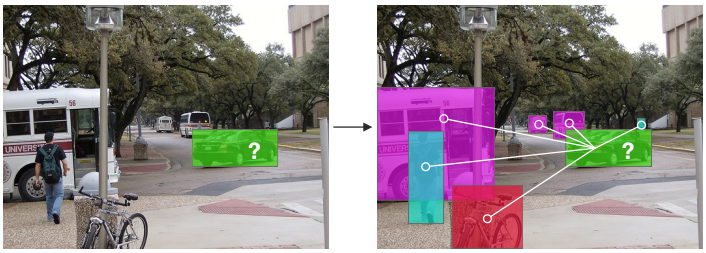


Figure 1. Current two-stage object detection frameworks classify each region in isolation, overlooking crucial contextual information (left). We propose a lightweight belief-propagation mechanism to bring contextual information to guide detection (right).

- Prior work tries to model context in a manner which is expensive from both computational and human labeling point of view [Chen et al., 2018, Jiang et al., 2018, Liu et al., 2018].
- We propose a novel approach for context-aware object detection by employing a lightweight belief-propagation mechanism which operates on visual representations of regions and the scene, as well as the spatial relationships between regions.
- We also experiment with capturing similarities between regions at a semantic level by modeling class co-occurrence and linguistic similarity between class names.

Approach

Our work builds on top of Structure Inference Net (SIN), proposed in Liu et al. [2018]. The first set of models we experiment with employ SIN's structure inference module as a post-processing step to bring semantic cues of different level for regions as in some prior work (e.g. Chen et al. [2018], Xu et al. [2019]). SIN uses two GRU cells (EdgeGRU and SceneGRU) as write functions for messages passed to a region from other regions as well as the scene. A message passed from region i to j is weighted as:

$$e_{i \to j} = \text{ReLU}(W_g R_{i \to j}) * \tanh(W_v [f^i, f^j])$$

In above formula, the former term captures spatial relationship between two regions while the latter captures their visual relationship. Please refer to our paper for the exact notations.

- **Base:** Models the co-occurrence of object categories based on the backbone detector's best set of guesses to capture the semantic relationship between regions for edge weight calculation.
- **Scene:** Updates scene representation at the end of each message-passing round, then uses this new representation in SceneGRU for the next round.
- **Attr1:** Models mid-level semantic relationships between regions using object category attributes for edge weight calculation. Having built an attribute matrix $\mathbb{A}^{C \times J}$, where C is the number of classes and J is the size of attribute dictionary, this model learns a projection $f : \mathbb{A}^{C \times J} \to \mathbb{M}^{C \times 16}$ to express attributes more complactly and then retrieves mid-level semantic similarity of regions from $\mathbb{M}\mathbb{M}^T$ using their highest predicted classes.
- **Attr2:** Similar to Attr1 but first maps attributes to regions using their predicted class scores. After this mapping, it projects region-attribute matrix to a lower dimensional space and retrieves attribute similarity of regions.

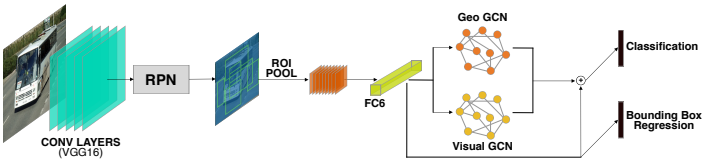


Figure 2. An overview of our most efficient approach. We model context between regions through single-layer GCNs which capture visual and spatial relationships.

Our second set of models replaces GRUs used in SIN with a lightweight belief-propagation mechanism.

- **GeoVis:** Employs two single-layer GCNs to perform message-passing between node pairs based on their visual (Visual GCN) and spatial (Geo GCN) relationships.
- **GeoVis-S:** Similar to GeoVis but models entire scene as a first-class participant in Visual GCN along with the regions.
- **GeoVis-Ling:** Uses a weighted loss formulation that seeks to increase the ability of a model to discriminate between semantically similar categories based on their pairwise distance in GloVe (Pennington et al. [2014]) space.

Experiments

	FRCNN	SIN	Scene	Attr1	Attr2	GeoVis-S	GeoVis-Ling
mAP	0.747	**0.765**	**0.756**	0.753	0.753	0.754	0.750
Animals AP	0.818	**0.825**	0.823	0.821	**0.825**	**0.829**	0.821

Table 1. Detection results on VOC 2007 test. VOC 2007 trainval + VOC 2012 trainval is used for training. The two best models are **bolded**. The top method is also **underlined**.

On VOC 2007, our Scene, Attr1 and Attr2 methods, and our GeoVis-S and GeoVis-Ling outperform FRCNN on most categories (please refer to our paper) and on average. **Importantly, our proposed method, GeoVis-S, outperforms SIN in terms of the average over animal categories, and it uses $6\times$ fewer parameters compared to SIN for context modeling.**

Test setting / Method	FRCNN	SIN	GeoVis-S
AP @[IoU=0.50:0.95 | area= all]	0.207	**0.215**	0.211
AP @[IoU=0.50 | area= all]	0.403	**0.423**	0.411
AP @[IoU=0.75 | area= all]	0.194	**0.198**	**0.198**

Table 2. Detection results on COCO 2019 test-dev (server evaluation). COCO 2014 train split is used for training. The best model is **bolded**.

On COCO 2019 test-dev, SIN improves over FRCNN by 4% but adds $12\times$ more parameters while our proposed method, GeoVis-S, brings a gain of 2% over FRCNN and adds $2\times$ more parameters. On the other hand, **GeoVis-S achieves the same performance as SIN when required IoU threshold is 0.75**.

On COCO 2014 minival (please refer to our paper), GeoVis-S achieves the best scores for 4 of the 11 supercategories while FRCNN being the best for only 1 supercategory.

Comparison of Model Parameters

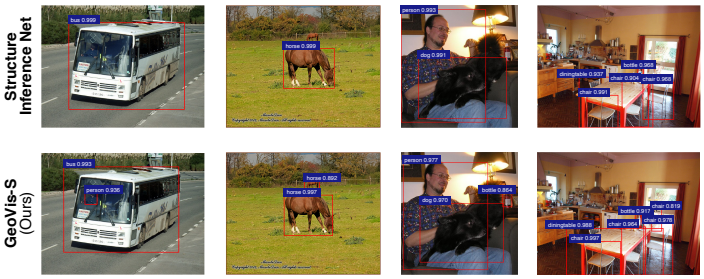
As all three models are identical up to FC6 and their R-CNN heads operate on $\mathbb{R}^{4096}$, we report the number of parameters between FC6 and the R-CNN head in each model to make a fair and dataset-agnostic comparison.

	FRCNN	SIN	GeoVis-S (Ours)
# Params	16,781,312	201,359,372	33,570,829

Table 3. Number of trainable parameters between FC6 and the R-CNN head.

Our model uses $6\times$ less parameters than SIN to model context, yet performs very competitively. This makes our model more feasible to deploy on resource-constrained devices.

Qualitative Evaluation



Above figure shows a qualitative comparison between SIN and our GEOVIS-S at 0.8 confidence threshold. As SIN passes messages between regions based on a single graphical representation wherein edges encode joint spatio-visual relationships between regions, it fails in utilizing context for rare object placements. In the first image it fails to detect the man who rides the bus since Pascal VOC contains very few examples with that particular spatio-visual relation between bus and person. Similarly, in the last image, SIN fails to detect the chair outside as it is under different illumination. Our method detects these two objects perfectly as it utilizes two graphs for message passing, separate for visual and spatial relationships, hence relaxes SIN's constraint.

References

Stephen E Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3:519–526, 1975.

Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7239–7248, 2018.

Chenhan Jiang, Hang Xu, Xiaodan Liang, and Liang Lin. Hybrid knowledge routed modules for large-scale object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1552–1563, 2018.

Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6985–6994, 2018.

Hang Xu, Chenhan Jiang, Xiaodan Liang, and Zhenguo Li. Spatial-aware graph relation network for large-scale object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9298–9307, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.