Unsupervised Learning of Visual Context from Instance Segmentation

Stella Yu, Tsung-Wei Ke (stellayu,twke@berkeley.edu) Alex Berg (aberg@fb.com)

Summary.

Real-world visual recognition is far more complex than object recognition: There is *stuff* without distinctive shape or appearance, and the same object appearing in different contexts calls for different actions. Yet, visual context is hard to describe and impossible to label manually. We consider visual context as semantic correlations between objects and their surroundings that include both object instances and stuff categories. We approach contextual object recognition as a pixel-wise feature representation learning problem that accomplishes supervised panoptic segmentation while discovering and encoding visual context automatically. Our experimental results on Cityscapes demonstrate that, in terms of surround semantics distributions, our retrievals are much more consistent with the query than the state-of-the-art segmentation method.

Related Works.

Instance contexts and relationships are explored mainly to enhance the detection performance. Earlier work [7] models the appearances and 2D spatial context as a graph. Hand-crafted features [8], tree-based models [3] are then developed to model co-occurring statistics and spatial configurations among object categories and object instances [10]. Recently, researchers integrate graphs [2] or spatial memory [1] into the deep learning framework. The distinction of our work is that our model does not explicitly model contexts yet is able to discovers novel contexts automatically. Our model is adapted from the Segment Sorting approach [5, 4, 6] for learning feature representations using supervised panoptic segmentation.

Research Approach.

We adapt the Segment Sorting approach [5] to panoptic segmentation by sorting segments according to both of its semantic and instance labels. With the learned feature representations, we classify segments into categories with a softmax classifier and merge them into instances by our proposed clustering algorithm.

We follow SegSort to formulate pixel-to-segment contrastive losses to enforce groupings and separations of pixel-wise embeddings. Since the loss formulation does not require a fixed number of classes as opposed to the conventional cross-entropy softmax loss, a way to extend it for instance discrimination is by changing the definition of ground truth labels and its corresponding selections of neighbor prototypes. We thus consider positive (negative) samples as any other 'same-instance' (all the other 'different-instance' and 'stuff-region') segments. Such trained embeddings group each instance against all the other instances, regardless of their semantic categories. We hypothesize that: **(1)** The embeddings encode the semantic labels inherently as instances of the same class appear similar. To extract such information, we stack a softmax classifier to predict the semantic class of each segment. **(2)** The embeddings encode object-centric context. This is endowed by the design of supervised semantic and instance segmentation with an unified representations. The feature of pedestrians walking across a road (on a sidewalk) encodes surrounding cars (buildings).

Given the panoptic embeddings and the resultant over-segmentations, the challenge is to group segments into instances correctly during inference. We need two criteria: 1) how to merge segments, and 2) when to stop the merging. To align with the formulation of the SegSort loss, we adopt a nearest neighbor clustering criterion [9] to greedily merge two segments with nearest prototypes, and stop the merging if the distance between two prototypes is greater than a selected threshold, or their dot product is less than the threshold.

Current Results on Contextual Retrievals on Cityscapes.