Weakly-supervised learning of visual relations

Julia Peyre¹,² ∙ Ivan Laptev¹,² ∙ Cordelia Schmid²
¹DI ENS, École normale supérieure, PSL Research University ∙ ²INRIA

**Goal**
- Detect visual relations using only image-level annotations
- Consider visual relations of the form (subject, predicate, object)

**Visual representation of a relation**
- Appearance features for individual objects from fc7 output of Fast-RCNN object detector
- Quantized spatial configuration between boxes (with GMM)

**Recall on Visual Relationship Dataset**
- **Metric**: recall@k, proportion of ground truth triplets retrieved among the top k detections

**Challenges**
- Diversity of visual relations
- Prohibitive cost of exhaustive manual annotation

**Contributions**
- Learning visual relations from image-level annotations
- UnRel: a new evaluation dataset with clean labels for evaluating the performance and generalization of relation detection

**Overview**
- Image-level triplets
  - person stand on surfboard
  - person carry person
  - person above surfboard
  - dog on surfboard
  - person taller than dog

**Visual of the spatial clusters**

**Learning with image-level labels**
- Discriminative clustering framework [1]:

  \[
  \min_{Z \in \mathcal{Z}} \min_{W \in \mathbb{R}^{d \times R}} \frac{1}{N} \left\| Z - XW \right\|_F^2 + \lambda \left\| W \right\|_F^2
  \]

  matching latent assignments
  regularizations

  Joint optimization of W and Z with Block-coordinate Frank-Wolfe algorithm [2]

**Retrieval for Unusual Relations (UnRel) Dataset**
- in the spirit of Out-of-Context dataset [1] ∙ 76 queries to test generalization
- 1071 images with box-level annotations for relations to evaluate in a clean setup
- 1533 relations annotated in total

**References**