CONTEXT-AWARE ZERO-SHOT RECOGNITION

Ruotian Luo — *TTI-Chicago* — rluo@ttic.edu

Ning Zhang — *Vaitl Inc.* — ning@vaitl.ai

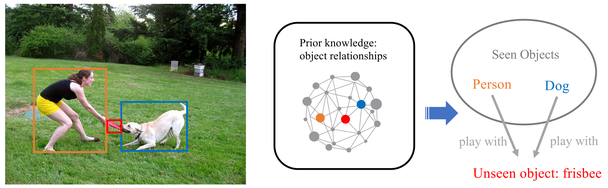
Bohyung Han — *Seoul National University* — bhhan@snu.ac.kr

Linjie Yang — *ByteDance* — yljatthu@gmail.com

Introduction

Traditional zero-shot learning which infers unseen categories by transferring knowledge from semantically similar seen categories, based on the assumption that similar semantic has similar appearance. The context information, the relations between objects is not explicitly used.

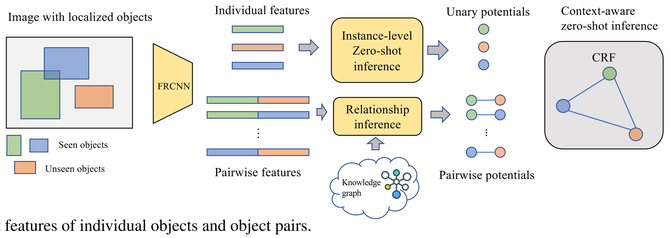
In this paper, we seek to infer novel objects in one image surrounded by other objects with the prior of object interactions. By leveraging the visual context and geometric relationships of all different objects in one image, we obtain information to infer unseen categories.



Problem formulation

Input: an image I and a set of regions $\{B_i\}$;
Output: class label $\{c_i\}$, $c_i \in \mathcal{C}$ for each region.
Class label space: full set $\mathcal{C}$; $\mathcal{C}$ can split to seen categories set $\mathcal{S}$ and unseen categories set $\mathcal{U}$.
During training time, only annotations of objects from $\mathcal{S}$ are provided. At test time, the model needs to classify regions of both seen and unseen categories. b
External knowledge graph indicates the possible interactions of a pair of object classes: $G = \{R_{mn}\}$, where R_{mn} is a collection of all the relations with class $\mathcal{C}_m$ as subject and class $\mathcal{C}_n$ as object. All relations set: $\mathcal{R}$, $R_{mn} \subseteq \mathcal{R}$.

Pipeline



1. Extract features of individual objects and object pairs.
2. Apply an instance-level zero-shot inference module on the individual features to generate a coarse probability of the object classes. Used as unary potential in the unified CRF model.
3. Apply relationship inference module takes pairwise features as input, integrated with knowledge graph and outputs the pairwise potentials.
4. MAP inference on the CRF model.

Our CRF inference model:

$$P(c_1 \dots c_N | B_1 \dots B_N) \propto \exp(\sum_i \theta(c_i|B_i) + \sum_{i \neq j} \phi(c_i, c_j|B_i, B_j))$$

Details

Unary potential
We use Fast-RCNN framework to get unary term.

$$\theta_i(c_i) = \log P_c(c_i|B_i) \tag{1}$$

Pairwise potential The relationship inference module takes a pair of regions as input, and outputs a relation potential vector: $F_r(r_k|B_i, B_j)$.
Each element $F_r(r|B_i, B_j)$ indicates how much the presence of relation r is supported by the region pair.
The pairwise potential of the CRF is formulated as:

$$\phi(c_i, c_j|B_i, B_j) = \sum_k \delta(r_k|c_i, c_j) F_r(r_k|B_i, B_j) \tag{2}$$

where $\delta(r_k|c_i, c_j)$ is an indicator function with value 1 otherwise 0 if relation $r_k \in R_{c_i,c_j}$. The pairwise potential is the sum of potentials of all possible relations between the pair of region labels. *Intuitively, a label assignment is encouraged when some relation r has high support given by F_r and also r is a potential relation between c_i and c_j.* $F_r(r|B_i, B_j)$ is a function of the relative geometry feature of two objects B_i and B_j.

$$F_r(r|B_i, B_j) = \text{MLP}(\mathcal{E}(g_{ij})) \tag{3}$$

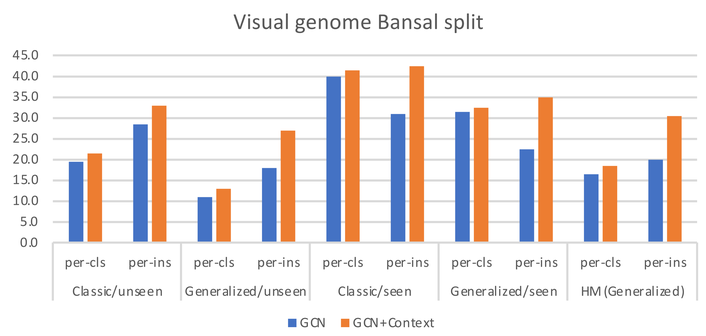
Training: Maximize pseudo-likelihood:

$$L = -\sum_i \log P(c_i^*|c_{\setminus i}^*) = -\sum_i \log \frac{\exp \sum_{j \neq i} [\theta_i(c_i^*) + \phi_{ij}(c_i^*, c_j^*) + \phi_{ji}(c_j^*, c_i^*)]}{\sum_c \exp \sum_{j \neq i} [\theta_i(c) + \phi_{ij}(c, c_j^*) + \phi_{ji}(c_j^*, c)]} \tag{4}$$

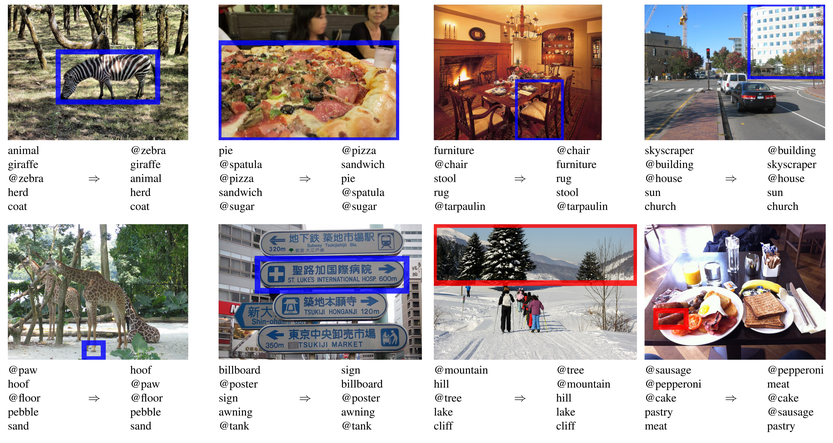
$F_r(r|B_i, B_j)$ is learned implicitly through optimizing of this loss. No ground truth annotation of relationships is used in training.

Quantitative results

Dataset
Visual Genome: 608 classes, 478 seen classes and 130 unseeen classes.
Knowledge graph is also extracted from Visual Genome. Only top 20 relations are kept in the graph.



Qualitative results



Examples of top-5 predictions change before(below left) and after(below right) context-aware inference. Blue boxes are examples of correct refinement and red ones denote failure cases. All unseen categories are prefixed with an @ for distinction.

Visualization of relation potentials. For a pair of objects, green box denotes subject and red one denotes object. The values of potential $F_r^{i,j}$ are shown on the right side of each image, respectively.

Visualization of pairwise potentials. Edges with potential less than 0.5 are omitted. The thickness of the line indicates how large the potential is. The ground truth category is annotated on the top-left corner of each box.

Conclusion

We design a novel framework to incorporate both instance-level and visual context knowledge to do zero-shot region recoginition task. Experiment results show that our context-aware approach is able to boost the performance by a large margin compared to models with only instance-level information. We believe that the new problem setting and the proposed algorithm will facilitate more interesting research for zero-shot and few-shot learning.