The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

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 AI Lab
linjie.yang@bytedance.com

Abstract

We present a novel problem setting in zero-shot learning, zero-shot object recognition and detection in the context. Contrary to the traditional zero-shot learning methods, which simply infers unseen categories by transferring knowledge from the objects belonging to semantically similar seen categories, we aim to understand the identity of the novel objects in an image surrounded by the known objects using the inter-object relation prior. Specifically, we leverage the visual context and the geometric relationships between all pairs of objects in a single image, and capture the information useful to infer unseen categories. We integrate our context-aware zero-shot learning framework into the traditional zero-shot learning techniques seamlessly using a Conditional Random Field (CRF). The proposed algorithm is evaluated on both zero-shot region classification and zero-shot detection tasks. The results on Visual Genome (VG) dataset show that our model significantly boosts performance with the additional visual context compared to traditional methods.

1 Introduction

Supervised object recognition has achieved substantial performance improvement thanks to the advance of deep convolutional neural networks in the last few years (Ren et al. 2015; Redmon and Farhadi 2017; Hu et al. 2018; Girshick 2015). Large-scale datasets with comprehensive annotations, *e.g.*, COCO (Lin et al. 2014), facilitate deep neural networks to learn semantic knowledge of the objects within a predefined set of classes.

However, it is impractical to obtain rich annotations for every class in the world while it is important to develop the models that can generalize to new categories without extra annotations. On the other hand, human beings have capability to understand the unseen object categories using external knowledge such as language descriptions and object relationships. The problem of inferring objects in unseen categories is referred to as *zero-shot* object recognition in recent literature (Fu et al. 2018; Xian et al. 2018).

In the absence of direct supervision, other resources of information such as semantic embedding (Norouzi et al. 2013), knowledge graph (Wang, Ye, and Gupta 2018; Rohrbach, Stark, and Schiele 2011), and attributes (Sung et al. 2018; Changpinyo et al. 2016) are often employed to infer the appearance of novel object categories through knowledge transfer from seen categories. The assumption behind the approaches is semantic similarity can transfer to visual similarity.

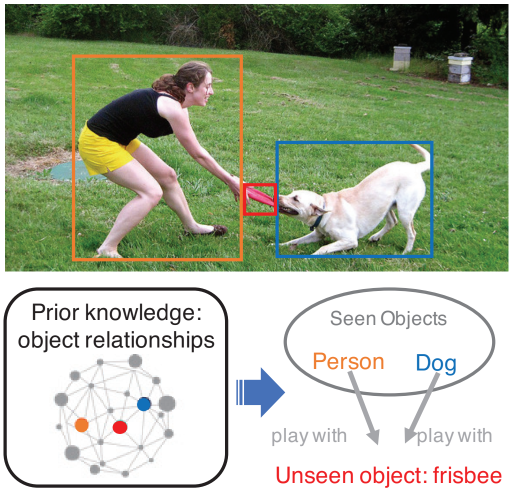

Figure 1: An example of zero-shot recognition with context information. It contains two seen objects (person and dog) and one unseen object (frisbee). The prior knowledge of relationships between seen and unseen categories provide cues to resolve the category of the unseen object.

Besides inferring novel object categories using visual similarity, human often capture the information of an object in the scene context. For example, if we do not know what the red disk-like object in Figure 1 is, it is possible to guess its category even with the limited visual cues by recognizing two other objects in the neighborhood, a person and a dog, using the prior knowledge that a person and a dog potentially play with together. Suppose that a frisbee is known to be such kind of an object, we can infer the object as a frisbee even without seeing it before. In this scenario, the interaction between multiple objects, *e.g.*. person, dog, and frisbee, provides additional clues to recognize the novel object—frisbee in this case; note that the external knowledge about the object relationships (person and dog can play with frisbee) is required for unseen object recognition.

Motivated by this intuition, we propose an algorithm for zero-shot image recognition *in the context*. Different from the traditional methods that infer each of unseen objects independently, we aim to recognize novel objects in the visual context, *i.e.*, by leveraging the relationships of the objects shown in an image. The relationship information is defined by a relationship knowledge graph in our framework and it is more straightforward to construct a knowledge graph than to collect dense annotations on images.

In our framework, a Conditional Random Field (CRF) is employed to jointly reason over local context information as well as relationship graph prior. Our algorithm is evaluated on Visual Genome dataset (Krishna et al. 2017), which provides a large number of object categories and diverse object relations; our model based on the proposed context knowledge representation illustrates the clear advantage when applied to various existing methods for zero-shot recognition. We believe the proposed topic will foster more interesting work in the domain of zero-shot recognition.

The main contributions of this work are as follows:

- We introduce a new framework of zero-shot learning in computer vision, referred to as zero-shot recognition in the context, where unseen object classes are identified by the relation to other ones shown in the same image.
- We propose a model for this task based on deep neural networks and CRF, which learns to leverage object relationship knowledge to recognize unseen object classes.
- The proposed algorithm achieves the significant improvement compared to existing methods on various models and settings that ignore visual context.

2 Related work

Zero-shot learning A wide range of external knowledge has been explored for zero-shot learning. Many use object attributes as a proxy to learn visual representation of unseen categories (Sung et al. 2018; Changpinyo et al. 2016; 2018). Semantic embeddings learned from large text corpus are also used to bridge seen and unseen categories (Frome et al. 2013; Norouzi et al. 2013). Combination of attributes and word embeddings are employed to learn classifiers of unseen categories by taking linear combinations of synthetic base classifiers (Changpinyo et al. 2016; 2018), and text descriptions are also incorporated later to predict classifier weights (Lei Ba et al. 2015). A recent work (Wang, Ye, and Gupta 2018; Kampffmeyer et al. 2018) applies Graph Convolutional Network (GCN) (Duvenaud et al. 2015) over WordNet knowledge graph to propagate classifier weights from seen to unseen categories. More detailed survey can be found in (Fu et al. 2018; Xian et al. 2018).

In addition to zero-shot recognition, zero-shot object detection (ZSD) task is also studied, which aims to localize individual objects of categories that are never seen during training (Bansal et al. 2018; Tao, Yang, and Cai 2018; Rahman, Khan, and Porikli 2018; Zhu et al. 2018; Demirel, Cinbis, and Ikizler-Cinbis 2018). Among the approaches, (Zhu et al. 2018) focuses on generating object proposals for unseen categories while (Bansal et al. 2018) trains a background-aware detector to alleviate the corruption of the "background" class with unseen classes. (Rahman, Khan, and Porikli 2018) proposes a novel loss function to reduce noise in semantic features. However, none of them have attempted to incorporate context information in the scene.

Context-aware detection Context information has been used to assist object detection before deep learning era (Galleguillos and Belongie 2010; Divvala et al. 2009; Felzenszwalb et al. 2010; Galleguillos, Rabinovich, and Belongie 2008; Desai, Ramanan, and Fowlkes 2011). Deep learning approaches such as Faster R-CNN (Ren et al. 2015) allow a region feature to look beyond its own bounding box via the large receptive field. Object relationships and visual context are also utilized to improve object detection. For example, Yang et al.; Li et al. show that the joint learning of scene graph generation and object detection improves detection results while Chen, Huang, and Tao; Hu et al. perform message passing between object proposals to refine detection results. A common-sense knowledge graph is used for weakly-supervised object detection (Kumar Singh et al. 2018).

Similar in motiation, Zablocki et al. recently take surrounding objects into consideration for zero-shot object recognition. Our method is complementary to theirs. However, we use geometric information along with a knowledge graph instead of only considering cooccurance in semantics.

Knowledge graphs Knowledge graphs has been applied to various vision tasks including image classification (Marino, Salakhutdinov, and Gupta 2016; Lee et al. 2018), zero-shot learning (Rohrbach et al. 2010; Rohrbach, Stark, and Schiele 2011; Wang, Ye, and Gupta 2018), visual reasoning (Malisiewicz and Efros 2009; Zhu, Fathi, and Fei-Fei 2014; Chen et al. 2018), and visual navigation (Yang et al. 2018b). Graph-based neural networks often propagate information on the knowledge graph (Marino, Salakhutdinov, and Gupta 2016; Lee et al. 2018; Wang, Ye, and Gupta 2018; Chen et al. 2018). Following (Marino, Salakhutdinov, and Gupta 2016; Chen et al. 2018; Yang et al. 2018b), we construct the relationship knowledge graph used in our method in a similar way.

3 Context-aware zero-shot recognition

3.1 Problem formulation

The existing zero-shot recognition techniquesmostly focus on classifying objects independently with no consideration of potentially interacting objects. To facilitate context-aware inference for zero-shot recognition, we propose to classify all the object instances—both seen and unseen objects—in an image. We first assume that the ground-truth bounding box annotations are given and propose to recognize objects in the unseen classes. After that, we also discuss zero-shot object detection when the ground-truth bounding boxes are not available at test time.

Our model takes an image I and a set of bounding boxes (regions) $\{B_i\}$ as its inputs, and produces a class label c_i out of the label set $\mathcal{C}$ for each region. Under the zero-shot recognition setting, the label set $\mathcal{C}$ is split into two disjoint subsets, $\mathcal{S}$ for seen categories and $\mathcal{U}$ for unseen categories. The object labels in $\mathcal{S}$ are available during training while the ones in $\mathcal{U}$ are not. The model needs to classify regions of both seen and unseen categories in testing.

Some existing zero-shot recognition approaches have utilized knowledge graph (Wang, Ye, and Gupta 2018) where the edges typically represent semantic similarity or hierarchy (*e.g.* zebra is equine). In our formulation, a relationship knowledge graph has edges representing the ordered pairwise relationships in the form of <subject, predicate, object> (*e.g.* <zebra, has, leg>), which indicate the possible interactions between a pair of objects in an image. A directed edge denotes a specific predicate (relation) in the relationship given by a tuple <subject, predicate, object>. We may have multiple relations for the same pair of categories; in other words, there can be multiple relationships defined on an ordered pair of categories. Given a set of relations, $\mathcal{R} = \{r_k | k = 1, \ldots, K\}$, the relationship graph is defined by $\mathsf{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ denotes a set of classes and $\mathcal{E} = \{r_{mn}^{(i)} \in \mathcal{R} | i = 1, \ldots, K_{mn} \text{ and } m, n \in \mathcal{C}\}$ is a set of directed edges representing relations between all pairs of a subject class m and an object class n. Note that K_{mn} is the number of all possible predicates between the ordered pair of classes.

3.2 Our framework

Our framework is illustrated in Figure 2. From an image with localized objects, we first extract features from the individual objects and the ordered object pairs. We then apply an instance-level zero-shot inference module to the individual object features, and obtain a probability distribution of the object over all object categories. The individual class likelihoods are used as unary potentials in the unified CRF model. A relationship inference module takes the pairwise features as an input and computes the corresponding pairwise potentials using the relationship graph.

Specifically, let B_i and c_i $(i = 1, \ldots, N)$ be an image region and a class assignment of N objects in an image. Our CRF inference model is given by

$$\begin{aligned} &P(c_1 \ldots c_N | B_1 \ldots B_N) \\ &\quad \propto \exp\left(\sum_i \theta(c_i|B_i) + \gamma \sum_{i \neq j} \phi(c_i, c_j | B_i, B_j)\right) \end{aligned} \tag{1}$$

where the unary potential $\theta(c_i|B_i)$ comes from the instance-level zero-shot inference module and the pairwise potential $\phi(c_i, c_j|B_i, B_j)$ is obtained from the relationship inference module. γ is a weight parameter balancing between unary and pairwise potentials.

The final prediction is generated through the MAP inference on the CRF model given by Eq. (1). We call the whole procedure *context-aware zero-shot inference*. Similar techniques can be found in context-aware object detection techniques (Divvala et al. 2009; Galleguillos and Belongie 2010). However, we claim that our algorithm has sufficient novelty because we introduce a new framework of zero-shot learning with context and design the unary and pairwise potentials specialized in CRF for zero-shot setting. We hereafter use $\theta_i(\cdot)$ and $\phi_{ij}(\cdot)$ as the abbreviations for $\theta(\cdot|B_i)$ and $\phi(\cdot|B_i, B_j)$, respectively. We discuss the detail of each component in the CRF next.

Instance-level zero-shot inference

We use a modified version of Fast R-CNN framework (Girshick 2015) to extract features from individual objects. The input image and the bounding boxes are passed through a network composed of convolutional layers and RoiAlign (He et al. 2017) layer. The network outputs a region feature $\mathbf{f}_i \in \mathbb{R}^{d_\mathbf{f}}$ for each region, which is further forwarded to a fully connected layer to produce the probability of each class $P_\mathrm{c}(c_i) = \mathrm{softmax}(\mathbf{W}\mathbf{f}_i)$, where $\mathbf{W} \in \mathbb{R}^{|\mathcal{C}| \times d_\mathbf{f}}$ is a weight matrix. The unary potential of the CRF is then given by

$$\theta_i(c_i) = \log P_c(c_i|B_i) \tag{2}$$

Although it is straightforward to learn the network parameters including $\mathbf{W}$ in the fully supervised setting, we can train the model only for the seen categories and obtain $\mathbf{W}_\mathcal{S} \in \mathbb{R}^{|\mathcal{S}| \times d_\mathbf{f}}$. To handle the classification of unseen category objects, we have to estimate $\mathbf{W}_\mathcal{U}$ as well and construct the full parameter matrix $\mathbf{W} = [\mathbf{W}_\mathcal{S}^\top, \mathbf{W}_\mathcal{U}^\top]^\top$ for prediction. There are several existing approaches (Changpinyo et al. 2016; Lei Ba et al. 2015; Changpinyo, Chao, and Sha 2017) to estimate the parameters for the unseen categories from external knowledge.

Relationship inference with relationship graph

The pairwise potential of the CRF model is given by a relationship inference module. It takes a pair of regions as its inputs and produces a relation potential, $\ell(\hat{r}_k; B_i, B_j)$, which indicates the likelihood of the relation $\hat{r}_k$ between the two bounding boxes. Then the pairwise potential of the CRF is formulated as

$$\phi(c_i, c_j | B_i, B_j) = \sum_k \delta(\hat{r}_k; c_i, c_j) \ell(\hat{r}_k; B_i, B_j), \tag{3}$$

where $\delta(\hat{r}_k; c_i, c_j)$ is an indicator function whether tuple $<c_i, \hat{r}_k, c_j>$ exists in the relationship graph. Intuitively, a label assignment is encouraged when the possible relations between the labels have large likelihoods.

The relationship inference module estimates the pairwise potential from a geometric configuration feature using an embedding function followed by a two-layer multilayer perceptron as

$$\ell(r|B_i, B_j) = \mathrm{MLP}(t_\eta(g_{ij})), \tag{4}$$

where g_{ij} is the relative geometry configuration feature of two objects corresponding to B_i and B_j based on (Hu et al. 2018) and $t_\eta(\cdot)$ embeds its input onto a high-dimensional space by computing cosine and sine functions of different wavelengths (Vaswani et al. 2017). Formally, translation- and scale-invariant feature g_{ij} is given by

$$g_{ij} = \left[\log \frac{|x_i - x_j|}{w_i}, \log \frac{|y_i - y_j|}{h_i}, \log \frac{w_j}{w_i}, \log \frac{h_j}{h_i}\right]^\top, \tag{5}$$

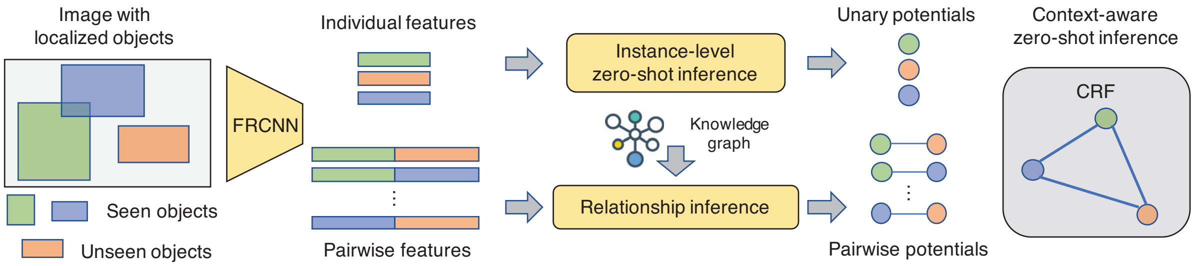

Figure 2: The overall pipeline of our algorithm. First, features for individual objects as well as object pairs are extracted from the image. An instance-level zero-shot inference module is applied on individual features to generate unary potentials. A relationship inference module takes pairwise features and relationship knowledge graph to generate pairwise potentials. Finally, the most likely object labels are inferred from CRF constructed by generated potentials.

where (x_i, y_i, w_i, h_i) represents the location and size of B_i.

To train the MLP in Eq. (4), we design a loss function based on pseudo-likelihood, which is the likelihood of a region given the ground-truth labels of the other regions. Maximizing the likelihood increases the potential of true label pairs while suppressing the wrong ones. Let c_i^* to be the ground-truth label of B_i. The training objective is to minimize the following loss function:

$$L = -\sum_i \log P(c_i^*|c_{\setminus i}^*), \qquad (6)$$

where $c_{\setminus i}^*$ denotes the ground-truth labels of bounding boxes other than B_i and

$$P(c_i^*|c_{\setminus i}^*) \qquad (7)$$

$$= \frac{\exp \sum_{j\neq i}[\theta_i(c_i^*) + \gamma\phi_{ij}(c_i^*, c_j^*) + \gamma\phi_{ji}(c_j^*, c_i^*)]}{\sum_{c\in\mathcal{S}} \exp \sum_{j\neq i}[\theta_i(c) + \gamma\phi_{ij}(c, c_j^*) + \gamma\phi_{ji}(c_j^*, c)]}.$$

Note that $\ell(r|B_i, B_j)$ is learned implicitly through optimizing of this loss. No ground-truth annotation about relationships is used in training.

Context-aware zero-shot inference The final step is to find the assignment that maximizes $P(c_1, \ldots, c_N)$ given the trained CRF defined by Eq. (1). We adopt mean field inference (Koller, Friedman, and Bach 2009) for efficient approximation. A distribution $Q(c_1, \ldots, c_N)$ is used to approximate $P(c_1, \ldots, c_N)$, which is given by the product of the independent marginals: $Q(c_1, \ldots, c_N) = \prod_i Q_i(c_i)$.

To get a good approximation of Q, we minimize the KL-divergence, $\mathrm{KL}(Q\|P)$. The optimal Q is obtained by iteratively updating Q using the following rule:

$$Q_i(c_i) \leftarrow \frac{1}{Z_i}\exp\left(\theta_i(c_i) + \gamma\sum_{j\neq i}\sum_{c_j\in\mathcal{C}} Q_j(c_j)\phi_{ij}(c_i, c_j)\right), \qquad (8)$$

where Z_i is a partition function. Then $\hat{c}_i = \mathrm{argmax}_{c_i} Q_i(c_i)$.

The pairwise potential defined in Eq. (3) involves a $(N \times |\mathcal{C}|)^2 \times |\mathcal{R}|$ matrix. Due to the huge computation/memory overhead when N and $|\mathcal{C}|$ are large, we perform pruning for acceleration, by selecting the categories with top K probabilities in terms of P_{c}. In this way, our method can be viewed as a cascade algorithm; the instance-level inference serves as the first layer of the cascade, and the context-aware inference refines the results using relationship information.

4 Implementation

Knowledge graph We extract our relationship knowledge graph from Visual Genome dataset, similar to (Marino, Salakhutdinov, and Gupta 2016).We first select 20 most frequent relations and collect all the subject-object relationships that (1) occurs more than 20 times in the dataset and (2) have the relation defined in $\mathcal{R}$. The purpose of this process is to obtain a knowledge graph with common relationships. The relation set $\mathcal{R}$ includes 'on', 'in', 'holding', 'wearing' etc.

Model We build our model based on a PyTorch Mask/Faster R-CNN (He et al. 2017) implementation[1], while the region proposal network and the bounding box regression branch are removed because ground-truth object regions are given. We use ResNet-50 (He et al. 2016) as our backbone model. Each image is resized with its shorter side 600 pixels.

Training We use a SGD with momentum to optimize all the modules. The instance-level zero-shot inference and relationship inference modules are trained separately in two stages. In the first stage, we train the instance-level zero-shot module on seen categories for 100K iterations. The model is fine-tuned from the pretrained ImageNet classification model. The learning rate is initialized to 0.005 and reduced by 10× after 60K and 80K iterations. After training on the seen categories, external algorithms are applied to transfer the knowledge to unseen categories. In the second stage, we train the relationship inference module for another 60k iterations with all the other modules fixed. To facilitate training, we omit unary potentials in Eq. (7) in practice. The learning rate is also initialized to 0.005 and reduced by 10× after 20K and 40K iterations. For all the modules, the parameter for the weight decay term is set to 0.0001, and the momentum is 0.9. The batch size is set to 8, and the batch normalization layers are fixed during training.

5 Experiments and results

5.1 Task

We mainly evaluate our system on zero shot region classification task. We provide ground-truth locations, $\{B_i\}$ for both training and testing. It enables us to decouple the recognition error from the mistakes from other modules including

[1] https://github.com/roytseng-tw/Detectron.pytorch

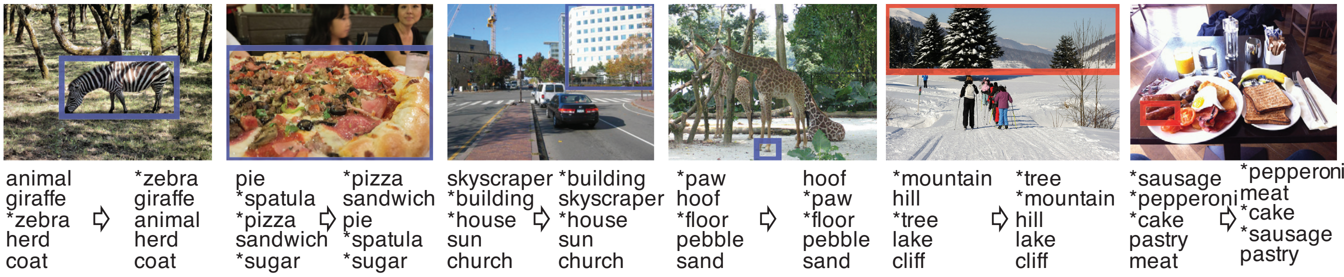

Figure 3: Examples of top-5 predictions change before (below left) and after (below right) context-aware inference. Blue boxes are examples of correct refinement and red ones denote failure cases. Each unseen category is prefixed with an * for distinction.

	Classic/unseen		Generalized/unseen		Classic/seen		Generalized/seen		HM (Generalized)	
	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins
WE	18.9	25.9	3.7	3.7	**35.6**	**57.9**	**33.8**	**56.1**	6.7	6.9
WE+Context	**19.5**	**28.5**	**4.1**	**10.0**	31.1	57.4	29.2	55.8	**7.2**	**17.0**
CONSE	**19.9**	27.7	0.1	0.6	**39.8**	31.7	**39.8**	31.7	0.2	1.2
CONSE+Context	19.6	**30.2**	**5.8**	**20.7**	29.6	**38.8**	25.7	**35.0**	**9.5**	**26.0**
GCN	19.5	28.2	11.0	18.0	39.9	31.0	31.3	22.4	16.3	20.0
GCN+Context	**21.2**	**33.1**	**12.7**	**26.7**	**41.3**	**42.4**	**32.2**	**35.0**	**18.2**	**30.3**
SYNC	25.8	33.6	12.4	17.0	39.9	31.0	34.2	24.4	18.2	20.0
SYNC+Context	**26.8**	**39.3**	**13.8**	**26.5**	**41.5**	**39.4**	**34.5**	**31.7**	**19.7**	**28.9**

Table 1: Results on Visual Genome dataset. Each group includes two rows. The upper one are baseline methods from zero-shot image classification literature. The lower ones are the results of their models attached with our context-aware inference. HM denotes harmonic mean of the accuracies on $\mathcal{S}$ and $\mathcal{U}$.

proposal generation, and diagnose clearly how much context helps zero-shot recognition on object level. As a natural extension of our work, we also evaluate on zero-shot detection task. In this case, we feed region proposals obtained from Edgeboxes (Zitnick and Dollár 2014) instead of ground-truth bounding boxes as input at test time.

5.2 Dataset

We evaluate our method on Visual Genome (VG) dataset (Krishna et al. 2017). VG contains two subsets of images, part-1 with ~60K images and part-2 with ~40K images. For our experiment, only a subset of categories are considered and the annotated relationships are not directly used.

We use the same seen and unseen category split in (Bansal et al. 2018). 608 categories are considered for classification. Among these, 478 are seen categories, and 130 are unseen categories. The part-1 of VG dataset are used for training, and randomly sampled images from part-2 are used for test. This results in 54,913 training images and 7,788 test images[2]. The relationship graph in this dataset has 6,396 edges.

5.3 Metrics and settings

We employ classification accuracy (AC) for evaluation, where results are aggregated in two ways; "per-class" computes the accuracy for each class and then computes the average over all classes while "per-instance" is the average accuracy over all regions. Intuitively, "per-class" metric gives more weight to the instances from rare classes than "per-instance" one.

The proposed algorithm is evaluated in both the classic and the generalized zero-shot settings. The model is only asked to predict among the unseen categories at test time in the classic setting while it needs to consider both seen and unseen categories under generalized setting. The generalized setting is more challenging than the classic setting because the model has to distinguish between seen and unseen categories.

5.4 Baseline methods

We compare our method with several baselines. Note that all baselines treat each object in an image as a separate image thus only utilizing instance-level features for inference.

Word Embedding (WE) In this method, weight vector $\mathbf{W}$ is set to be the GloVe (Pennington, Socher, and Manning 2014) word embedding of each category. $\mathbf{W}$ is fixed during training. (Note that the same word embedding is used for the other settings.)

CONSE (Norouzi et al. 2013) CONSE first trains classifiers on $\mathcal{S}$ with full supervision. At test time, each instance in an unseen class is embedded onto the word embedding space by a weighted sum of the seen category embeddings, where the weights are given by the classifier defined on $\mathcal{S}$. Then the image is predicted to the closest unseen (and seen in the

[2] The training images still include instances of unseen categories, because pure images with only seen categories are too few. However, we only use annotations of seen categories.

	Classic/unseen		Generalized/unseen		Classic/seen		Generalized/seen	
	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins
GCN	19.5	28.2	11.0	18.0	39.9	31.0	31.3	22.4
GCN+G	**21.2**	**33.1**	**12.7**	**26.7**	**41.3**	42.4	32.2	35.0
GCN+GA	20.4	26.5	9.2	15.3	40.9	**44.8**	**34.7**	**40.9**
SYNC	25.8	33.6	12.4	17.0	39.9	31.0	34.2	24.4
SYNC+G	**26.8**	**39.3**	**13.8**	**26.5**	41.5	39.4	34.5	31.7
SYNC+GA	26.6	33.6	11.3	16.4	**41.6**	**42.8**	**36.5**	**38.5**

Table 2: Results of different inputs to relationship inference module. *+G is the model with only geometry information. *+GA is the model with both geometry and appearance feature.

	Generalized		Classic	
	top-1	top-5	top-1	top-5
WE+Ctx	3.7→10.0	26.6	25.9→28.5	57.5
CONSE+Ctx	0.6→20.7	29.4	27.7→30.2	56.1
GCN+Ctx	18.0→26.7	38.3	28.2→33.1	51.6
SYNC+Ctx	17.0→26.5	49.4	33.6→39.3	68.9

Table 3: Per-instance top-K accuracy on unseen categories.

generalized setting) class in the word embedding space.

GCN (Wang, Ye, and Gupta 2018) Similar to CONSE, GCN first trains classifiers on $\mathcal{S}$. Then it learns a GCN model to predict classifier weights for $\mathcal{U}$ from the model for the seen classes. The GCN takes the word embeddings of all the seen and unseen categories and the classifier weights of $\mathcal{S}$ as its inputs, and learns the global classifier weights by regression. In the end, the predicted classifier weights are used in the inference module for both seen and unseen categories. We use a two-layer GCN with LeakyReLU as the activation function. Following (Wang, Ye, and Gupta 2018), we use WordNet (Miller 1995) to build the graph.

SYNC (Changpinyo et al. 2016; 2018) This approach aligns semantic and visual manifolds via use of *phantom* classes. The weight of phantom classifier is trained to minimize the distortion error as $\min_{\mathbf{V}} \|\mathbf{W}_{\mathcal{S}} - \mathbf{S}_{\mathcal{S}}\mathbf{V}\|$, where $\mathbf{S}_{\mathcal{S}}$ is the semantic similarity matrix between seen categories and phantom classes and $\mathbf{V}$ is the model parameter of the phantom classifier. The classifier weights for $\mathcal{U}$ is given by a convex combinations of phantom classifier as $\mathbf{W}_{\mathcal{U}} = \mathbf{S}_{\mathcal{U}}\mathbf{V}$, where $\mathbf{S}_{\mathcal{U}}$ is the semantic similarity matrix between unseen categories and phantom classes.

5.5 Zero-shot recognition results

Table 1 presents the performance of our context-aware algorithm based on the four zero-shot recognition baseline methods. On all backbone baselines, our model improves the accuracy on both unseen categories, both in classic and generalized settings. The performances on seen categories are less consistent, which is mainly due to the characteristics of baseline methods, but still better in general.

For WE and CONSE methods, we can see that there are huge accuracy gaps between seen and unseen categories, especially under generalized setting. This implies that the backbone models are biased towards seen categories significantly. Hence, it is natural that our model sacrifices accuracy on $\mathcal{S}$ to improve performance on $\mathcal{U}$. GCN and SYNC, on the contrary, are more balanced, and our algorithm is able to consistently improve on both seen and unseen categories combined with GCN and SYNC.

The harmonic means of accuracies on seen and unseen categories are consistently higher in our method than in the baselines under generalized setting. Note that this metric is effective to compare overall performance on both seen and unseen categories as suggested in (Xian et al. 2018).

Top-K refinement As we mentioned, our pruning method makes the context-aware inference a top-k class reranking. We conduct current experiment with $K = 5$. In Table 3, we show "per-instance" top-1 accuracy versus top-5 accuracy of different algorithms on unseen categories. Since we only rerank the top-5 classes, the top-5 accuracies do not change, and the top-1 accuracy are upper bounded by the corresponding top-5 accuracy. After applying context-aware inference, the top-1 accuracies increase. Notably, the baseline model of CONSE has near 0 accuracy under generalized setting because it biases towards seen categories severely. However, its top-5 accuracy is reasonable. Our method is able to reevaluate top-5 predictions with the help of relation knowledge and increase the top-1 accuracy significantly.

Qualitative results Figure 3 shows qualitative results from the context-aware inference. Our context-aware model adjusts the class probabilities based on the object context. For example, zebra is promoted in the first image because the bands on its body while sausage helps recognize the pizza in the second image. Different patterns can be found for label refinement: general to specific (animal to zebra), specific to general (skyscraper to building), and corrected to similar objects (pie to pizza, paw to hoof).

Input choices for relationship inference Our relationship inference module only takes geometry information as input to avoid overfitting to seen categories. One alternative we tried is combining it with region appearance feature. We project region features f_i and f_j into lower dimension and concatenate it with $\mathcal{E}(g_{ij})$ to produce relation potentials. We report the results in Table 2. The appearance augmented relationship inference module is named as +GA in the table. It's shown that +GA biases towards seen categories, and hurts performance on unseen categories. +GA on generalized setting on unseen categories is even worse than the baselines.

	Unseen		Seen		HM	
	0.4	0.5	0.4	0.5	0.4	0.5
GCN	8.5	6.2	**23.1**	**17.8**	12.4	9.2
GCN+Ctx	**9.7**	**6.9**	22.3	16.0	**13.5**	**9.6**
SYNC	11.1	8.2	**24.2**	**18.8**	15.2	11.4
SYNC+Ctx	**12.0**	**8.6**	23.1	17.4	**15.8**	**11.5**

Table 4: Generalized zero-shot detection results. Recall@100 with IOU threshold 0.4/0.5 is reported. HM denotes harmonic mean.

5.6 Zero-shot detection results

We extend our region classification model for detection task by adding a background detector. We set the classifier weight of background class to be normalized average classifier weights: $\mathbf{W}_{\text{bg}} = \frac{\sum_{c\in\mathcal{C}}\mathbf{W}_c}{\|\sum_{c\in\mathcal{C}}\mathbf{W}_c\|^2}$, where each row of $\mathbf{W}$ needs to be normalized in advance. Furthermore, given thousands of region proposals, we only consider the top 100 boxes with highest class scores given by instance-level module for context-aware inference.

Following (Bansal et al. 2018), EdgeBoxes proposals are extracted for test images, where only proposals with scores higher than 0.07 are selected. After detection, non-maximum suppression is applied with IOU threshold 0.4. Due to incomplete annotations in VG, we report Recall@100 scores with IOU threshold 0.4/0.5. Table 4 presents instance-level zero-shot performance of GCN and SYNC models, where our method shows improved accuracy on unseen categories and higher overall recalls given by harmonic means. Note that our results on the generalized zero-shot setting already outperforms the results on the classic setting reported in (Bansal et al. 2018).

6 Conclusions

We presented a novel setting for zero-shot object recognition, where high-level visual context information is employed for inference. Under this setting, we proposed a novel algorithm to incorporate both instance-level and object relationship knowledge in a principled way. Experimental results show that our approach boosts the performance compared to the models with only instance-level information. We believe that this new problem setting and the proposed algorithm facilitate more interesting research for zero-shot or few-shot learning.

Acknowledgements. This research was partly supported by the Korean ICT R&D program of the MSIP/IITP grant [2016-0-00563, 2017-0-01778], the Air Force Research Laboratory and DARPA under awards number FA8750-18-2-0126 and FA9550-18-1-0166.

Appendices

ImageNet overlaps $\mathcal{U}$ 12 classes in the $\mathcal{U}$ of (Bansal et al. 2018) are also in ImageNet1k. To get true zero-shot performance, in Table 5 we report the per-class accuracy on the remaining 118 unseen categories. Results show that our methods still outperform baselines.

	Classic	Generalized
GCN	18.3	10.9
GCN+Context	**20.1**	**12.5**
SYNC	25.2	12.2
SYNC+Context	**26.1**	**13.7**

Table 5: Per-class accuracy after removing ImageNet categories in unseen set.

Choice of γ γ is chosen by cross validation. For WE, CONSE, GCN, γ is 1, and for SYNC, γ is set to be 0.5.

References

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-shot object detection. In *The European Conference on Computer Vision (ECCV)*.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5327–5336.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2018. Classifier and exemplar synthesis for zero-shot learning. *arXiv preprint arXiv:1812.06423*.

Changpinyo, S.; Chao, W.-L.; and Sha, F. 2017. Predicting visual exemplars of unseen classes for zero-shot learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 3496–3505. IEEE.

Chen, X.; Li, L.-J.; Fei-Fei, L.; and Gupta, A. 2018. Iterative visual reasoning beyond convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Z.; Huang, S.; and Tao, D. 2018. Context refinement for object detection. In *The European Conference on Computer Vision (ECCV)*.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2018. Zero-shot object detection by hybrid region embedding. *arXiv preprint arXiv:1805.06157*.

Desai, C.; Ramanan, D.; and Fowlkes, C. C. 2011. Discriminative models for multi-class object layout. *International journal of computer vision* 95(1):1–12.

Divvala, S. K.; Hoiem, D.; Hays, J. H.; Efros, A. A.; and Hebert, M. 2009. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1271–1278. IEEE.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32(9):1627–1645.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Fu, Y.; Xiang, T.; Jiang, Y.-G.; Xue, X.; Sigal, L.; and Gong, S. 2018. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine* 35(1):112–125.

Galleguillos, C., and Belongie, S. 2010. Context based object categorization: A critical survey. *Computer vision and image understanding* 114(6):712–722.

Galleguillos, C.; Rabinovich, A.; and Belongie, S. 2008. Object categorization using co-occurrence, location and appearance. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2980–2988. IEEE.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kampffmeyer, M.; Chen, Y.; Liang, X.; Wang, H.; Zhang, Y.; and Xing, E. P. 2018. Rethinking knowledge graph propagation for zero-shot learning. *arXiv preprint arXiv:1805.11724*.

Koller, D.; Friedman, N.; and Bach, F. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Kumar Singh, K.; Divvala, S.; Farhadi, A.; and Jae Lee, Y. 2018. Dock: Detecting objects by transferring common-sense knowledge. In *The European Conference on Computer Vision (ECCV)*.

Lee, C.-W.; Fang, W.; Yeh, C.-K.; and Wang, Y.-C. F. 2018. Multi-label zero-shot learning with structured knowledge graphs.

Lei Ba, J.; Swersky, K.; Fidler, S.; et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 4247–4255.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Malisiewicz, T., and Efros, A. 2009. Beyond categories: The visual memex model for reasoning about object relationships. In *Advances in neural information processing systems*.

Marino, K.; Salakhutdinov, R.; and Gupta, A. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Rahman, S.; Khan, S.; and Porikli, F. 2018. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *arXiv preprint arXiv:1803.06049*.

Redmon, J., and Farhadi, A. 2017. Yolo9000: better, faster, stronger. *arXiv preprint*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Rohrbach, M.; Stark, M.; Szarvas, G.; Gurevych, I.; and Schiele, B. 2010. What helps where–and why? semantic relatedness for knowledge transfer. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 910–917. IEEE.

Rohrbach, M.; Stark, M.; and Schiele, B. 2011. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1641–1648. IEEE.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tao, Q.; Yang, H.; and Cai, J. 2018. Zero-annotation object detection with web knowledge transfer. In *The European Conference on Computer Vision (ECCV)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6857–6866.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*.

Yang, J.; Lu, J.; Lee, S.; Batra, D.; and Parikh, D. 2018a. Graph r-cnn for scene graph generation. In *The European Conference on Computer Vision (ECCV)*.

Yang, W.; Wang, X.; Farhadi, A.; Gupta, A.; and Mottaghi, R. 2018b. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*.

Zablocki, E.; Bordes, P.; Piwowarski, B.; Soulier, L.; and Gallinari, P. 2019. Context-aware zero-shot learning for object recognition. *arXiv preprint arXiv:1904.12638*.

Zhu, P.; Wang, H.; Bolukbasi, T.; and Saligrama, V. 2018. Zero-shot detection. *arXiv preprint arXiv:1803.07113*.

Zhu, Y.; Fathi, A.; and Fei-Fei, L. 2014. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*.

Zitnick, C. L., and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*.