

Context-Aware Zero-Shot Recognition

*Ruotian Luo, Ning Zhang,
Bohyung Han, Linjie Yang;*
arXiv, 2019

Context-Aware Zero-Shot Recognition



??

Context-Aware Zero-Shot Recognition



꽃인가?

Context-Aware Zero-Shot Recognition

- Intro.

- 기존 ZSL 기법은 간단하게 지식을 transfe해서 단일 물체에 대해 분류
- ‘의미적으로 가까운 물체는 보이기에다 비슷할거다’ 라는 가정
- 하지만 비슷하게 생겼지만 전혀 무관한 물체도 많음
(ex. 접시 vs 원판, 미러볼 vs 축구공)
- 보통 사람은 처음 보는 물체도 전체 장면을 보고 무엇인지 추정을 함
- 따라서 본 논문에서는 주변 상황정보(context)를 반영하여 분류 정확도를 높이하고자 함

Context-Aware Zero-Shot Recognition

- Intro.

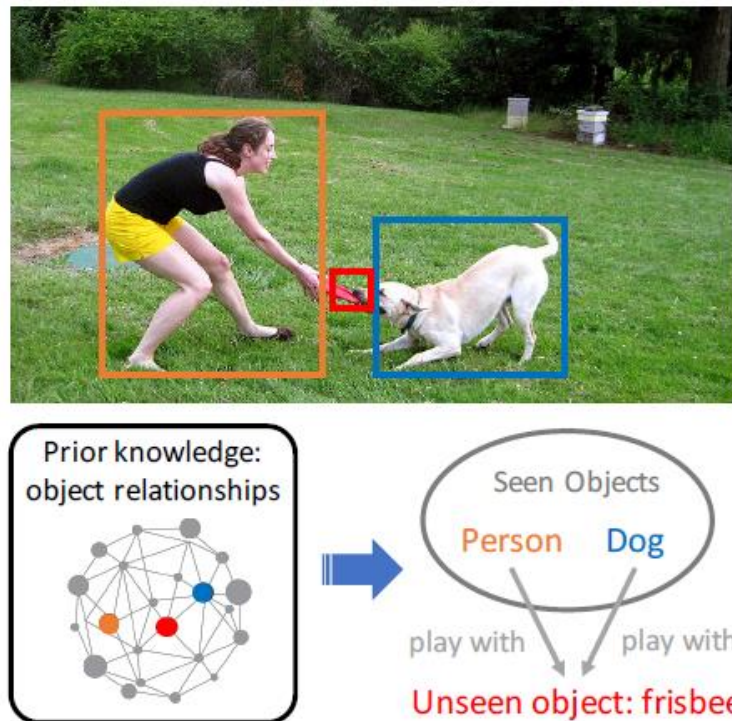
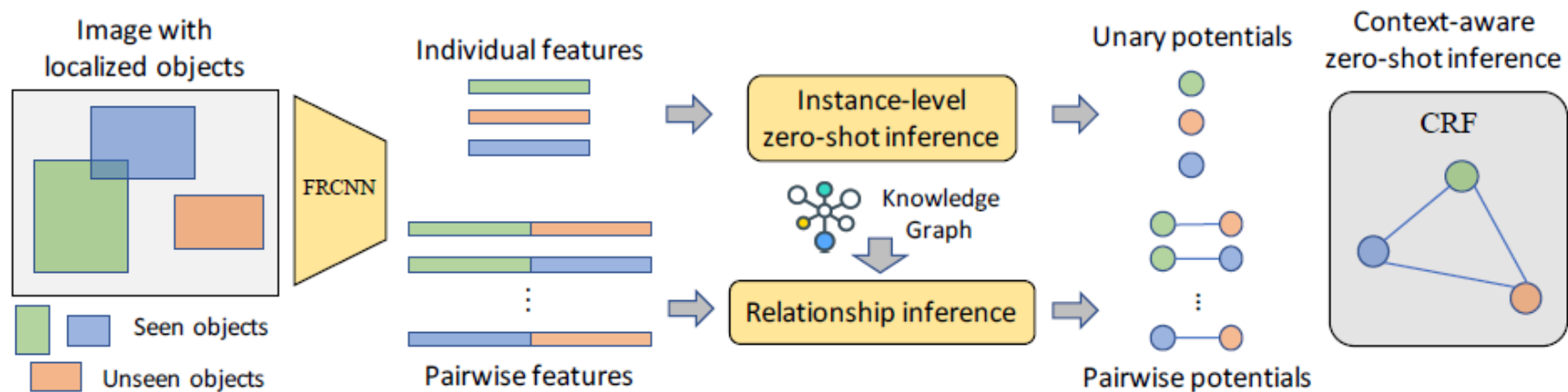


Figure 1. An example of zero-shot recognition with context information. It contains two seen objects (person and dog) and one unseen object (frisbee). The prior knowledge of relationships between seen and unseen categories provide cues to resolve the category of the unseen object.

Context-Aware Zero-Shot Recognition

Method



Context-Aware Zero-Shot Recognition

- Method

- based on CRF model

$$P(c_1 \dots c_N | B_1 \dots B_N)$$

$$\propto \exp \left(\sum_i \theta(c_i | B_i) + \gamma \sum_{i \neq j} \phi(c_i, c_j | B_i, B_j) \right)$$

Potential functions

c: class

B: bounding box (image)

γ : scaling parameter

Context-Aware Zero-Shot Recognition

- Method

- based on CRF model

$$P(c_1 \dots c_N | B_1 \dots B_N) \\ \propto \exp \left(\underbrace{\sum_i \theta(c_i | B_i)}_{\text{unary potential}} + \gamma \underbrace{\sum_{i \neq j} \phi(c_i, c_j | B_i, B_j)}_{\text{pairwise potential}} \right)$$

Context-Aware Zero-Shot Recognition

▪ Method

- based on CRF model

$$P(c_1 \dots c_N | B_1 \dots B_N) \\ \propto \exp \left(\underbrace{\sum_i \theta(c_i | B_i)}_{\text{unary potential}} + \gamma \sum_{i \neq j} \phi(c_i, c_j | B_i, B_j) \right)$$

$$\theta_i(c_i) = \log P_c(c_i | B_i)$$

f : region feature

$$P_c(c_i) = \text{softmax}(\bar{W}f_i)$$

W : weight matrix (classifier)

- W 는 추상적인 함수로, 실제로는 다른 기존 ZSL 기법들의 분류 결과를 활용

Context-Aware Zero-Shot Recognition

▪ Method

- based on CRF model

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

pairwise potential

$$\phi(c_i, c_j | B_i, B_j) = \sum_k \underset{1}{\delta(\hat{r}_k; c_i, c_j)} \underset{2}{\ell(\hat{r}_k; B_i, B_j)}$$

r : relation

1 : 두 class 사이에 해당 relation이 있나 (indicator function)

2 : 두 bounding box로부터 해당 relation이 추정될 likelihood

Context-Aware Zero-Shot Recognition

▪ Method

- based on CRF model

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

pairwise potential

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

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

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

- Bounding box의 위치 정보로 relation의 likelihood 산출

Context-Aware Zero-Shot Recognition

▪ Method

- based on CRF model

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

pairwise potential

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

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

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

- t_n 은 positional encoding 후 g랑 summation 하는 함수쓰로 예상됨

Context-Aware Zero-Shot Recognition

▪ Method

- based on CRF model

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

pairwise potential

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

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

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

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

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

(7)

- increases the potential of true label pairs

Context-Aware Zero-Shot Recognition

- Implementation

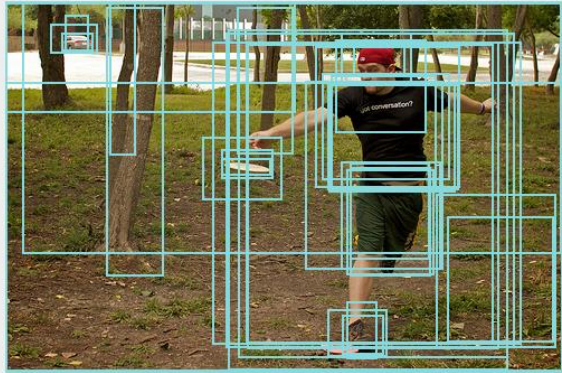
- 생략

Context-Aware Zero-Shot Recognition

▪ Experiment

- Visual Genome dataset (multi-object imageset)으로만 학습하고 실험 진행
- Part-1 VG 는 training, Part-2 VG는 test에 활용
- Visual Genome에 이미 물체 간의 관계 정보도 포함되어 있음

Regions	Attributes	Relationships
Man throwing frisbee	cap is red	tree IN urban park
White frisbee in midair	tree trunk is thin	man WEARING green shorts
Red baseball cap worn backwards	tree trunk is straight	man WEARING black shirt
Thin straight tree trunk	grass is patchy	man WEARING sneakers
Patchy park grass	grass is dark	guy WEARING clothes
Black tee shirt printed with text	shorts is dark	man throwing frisbee
Dark green shorts	shorts is green	man aiming and throwing frisbee
Parked grey car	car is grey	man IN field
Trees in an urban	car is parked	
	green shorts is green	
	black shirt is black	



²The 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.

Context-Aware Zero-Shot Recognition

▪ Experiment

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} .

	Classic/unseen		Generalized/unseen		Classic/seen		Generalized/seen		HM (Generalized)	
	per-cl	per-ins	per-cl	per-ins	per-cl	per-ins	per-cl	per-ins	per-cl	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

Context-Aware Zero-Shot Recognition

- Experiment

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.

	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

- G : 제시했던 방법

- GA : Bounding box 위치 정보와 image feature를 concat해서 사용해 봄
(결론, seen image에 overfitting 되는 경향)

Context-Aware Zero-Shot Recognition

- Experiment



animal		@zebra
giraffe		giraffe
@zebra	⇒	animal
herd		herd
coat		coat



pie		@pizza
@spatula		sandwich
@pizza	⇒	pie
sandwich		@spatula
@sugar		@sugar



furniture		@chair
@chair		furniture
stool	⇒	rug
rug		stool
@tarpaulin		@tarpaulin



skyscraper		@building
@building		skyscraper
@house	⇒	@house
sun		sun
church		church