https://arxiv.org/abs/1904.03181

# Detecting Human-Object Interactions via Functional Generalization

Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, Rama Chellappa
University of Maryland, College Park

AAAI 2020 예정

인공지능 연구실
석사과정 구자봉

# 문제 정의 :
# HOI(Human-object interaction)

⟨human, ride, bicycle⟩
⟨human, sit_on, bicycle⟩
⟨human, straddle, bicycle⟩

# 기존 방법의 문제점 제시 :
# 인간과 오브젝트 쌍 사이에서 가능한 관계의 개수가 너무 많다

HICO-DET
Images 47,776 (38,118, 9,658)
Objects 80 (airplane, apple…)
Verbs 117 (carry, catch…)
HOI 600 (airplane – board, direct, exit, fly…)

실제 가능한 관계 개수 :
80 (Objects) X 117 (Verbs) = 9360 (interactions)

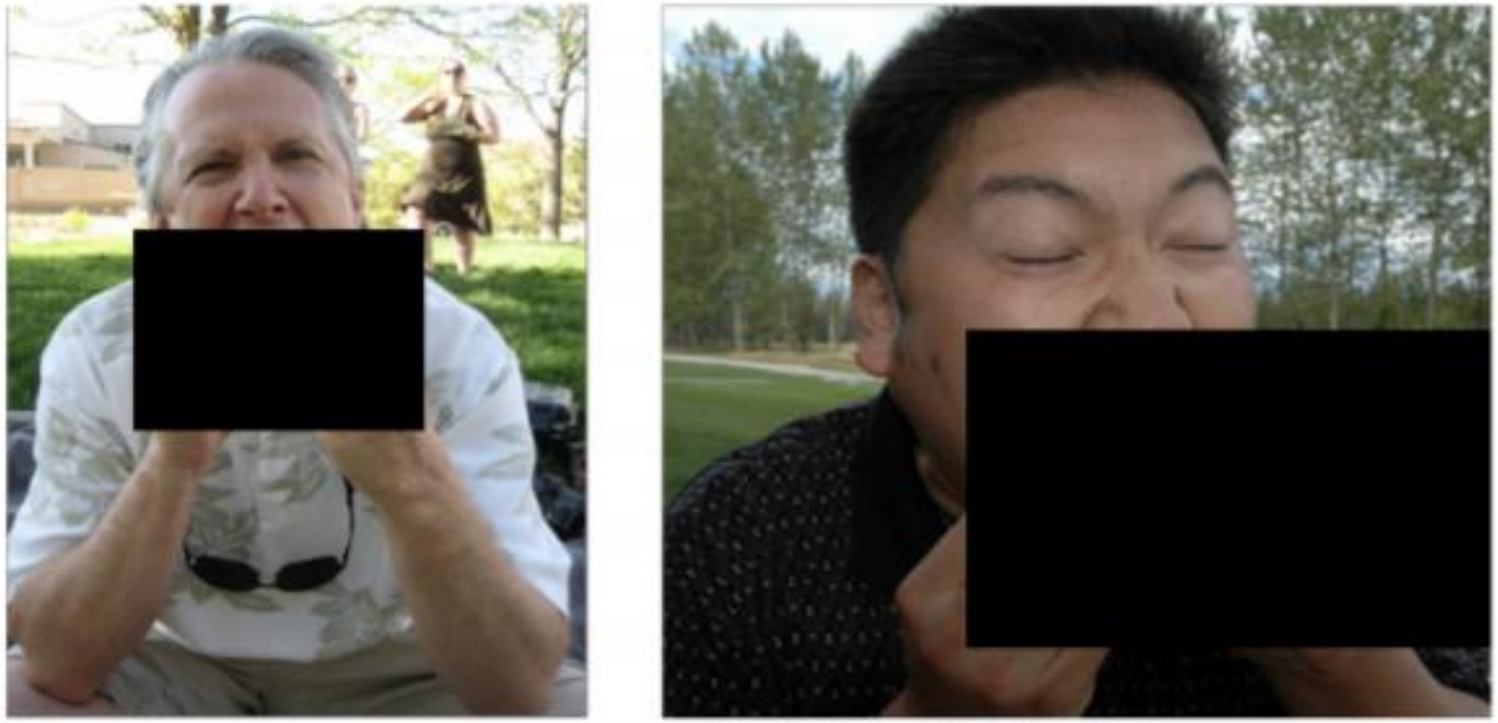
But 600개로 제한하여 데이터를 생성함

이는 데이터 상에서는 잘 되지만 일반화 시켰다고는 할 수 없음

⟨human, push, car⟩

# 동기 :
# 인간은 비슷한 방식으로 기능적으로 유사한 물체와 상호작용을 한다.

(zero-shot recognition)



https://perpetual.tistory.com/76

# Zero shot learning(제로샷 학습)

## 기본 정의: 한 번도 듣도보도 못했던 클래스를 분류하도록 학습하는 것

기존에 승용차라는 데이터가 많아서 승용차 클래스에 대해서 성공적으로 학습을 했음.
근데 입력이 갑자기 트럭이 갑자기 나타남.

바람컨데 그럼 알아서 트럭이라는 승용차와 가까운 클래스를 짜잔하고 만들어주면 좋겠지만,
기존 모델들은 안습적이게도 일반적으로 노답이 됨 ㅠㅠㅠ
(정확히는 기존 클래스에 분류해버림)

그러한 상황을 해결하기 위해 만든 방법이 zero-shot learning
알아서 새로운 클래스를 뚝딱 만들어서 첨보는 트럭을 분류해주는 식으로 말이다

## Sparse Representation : 희소 표현(클래스 수==차원수), 원 핫 벡터

Ex) 강아지 = [ 0 0 0 0 1 0 0 0 0 0 0 0 ... 중략 ... 0] # 이 때 1 뒤의 0의 수는 9995개.

## Dense Representation : 밀집 표현(사용자 임의의 차원 정함), 워드 임베딩, 임베딩 벡터

Ex) 강아지 = [0.2 1.8 1.1 -2.1 1.1 2.8 ... 중략 ...] # 이 벡터의 차원은 128

| - | 원-핫 벡터 | 임베딩 벡터 |
|---|---|---|
| 차원 | 고차원(단어 집합의 크기) | 저차원 |
| 다른 표현 | 희소 벡터의 일종 | 밀집 벡터의 일종 |
| 표현 방법 | 수동 | 훈련 데이터로부터 학습함 |
| 값의 타입 | 1과 0 | 실수 |

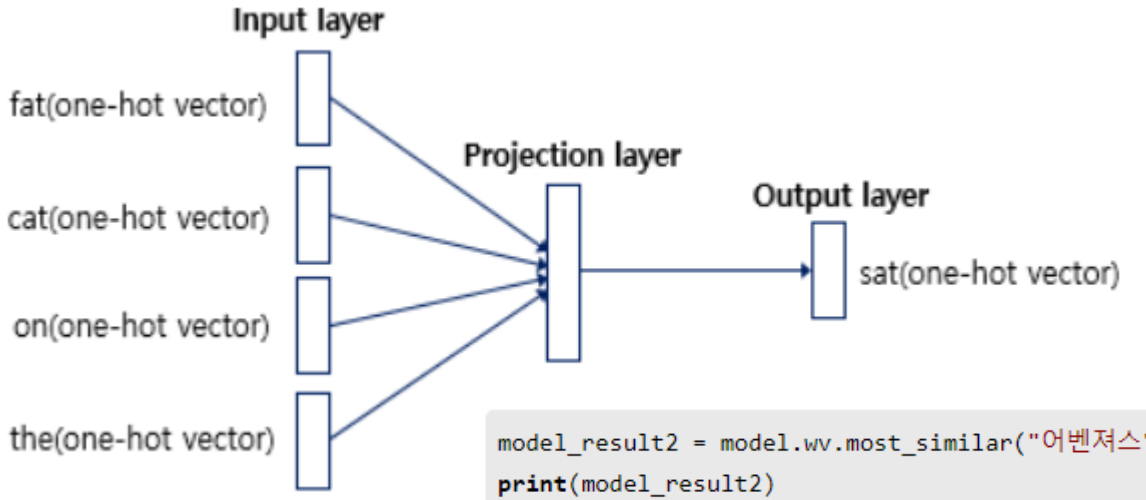
**distributed representation : 분산 표현, 단어의 의미를 다차원 공간에 벡터화, 임베딩벡터, 밀직벡터에도 속함**
**비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다**

Ex) 강아지 = [0.2 0.3 0.5 0.7 0.2 ... 중략 ... 0.2]

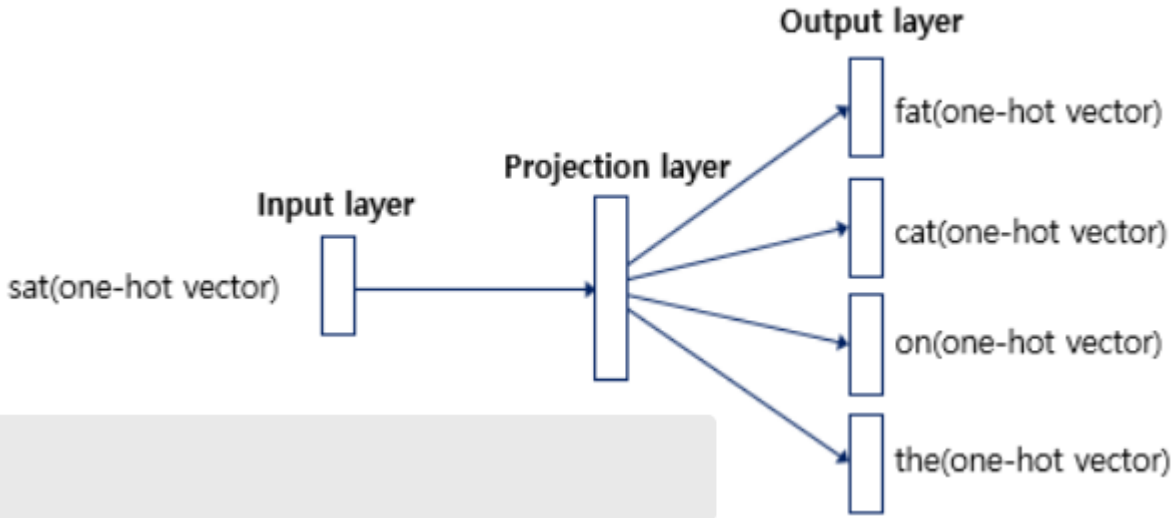
**Word2Vec : 워드 투 벡터, 단어 간의 유사도를 반영할 수 있도록 단어를 분산 표현으로 만드는 방법**

**예문 : "The fat cat sat on the mat"**

**CBOW(Continuous Bag of Words) :**
**주위 단어로 중간의 단어 예측**



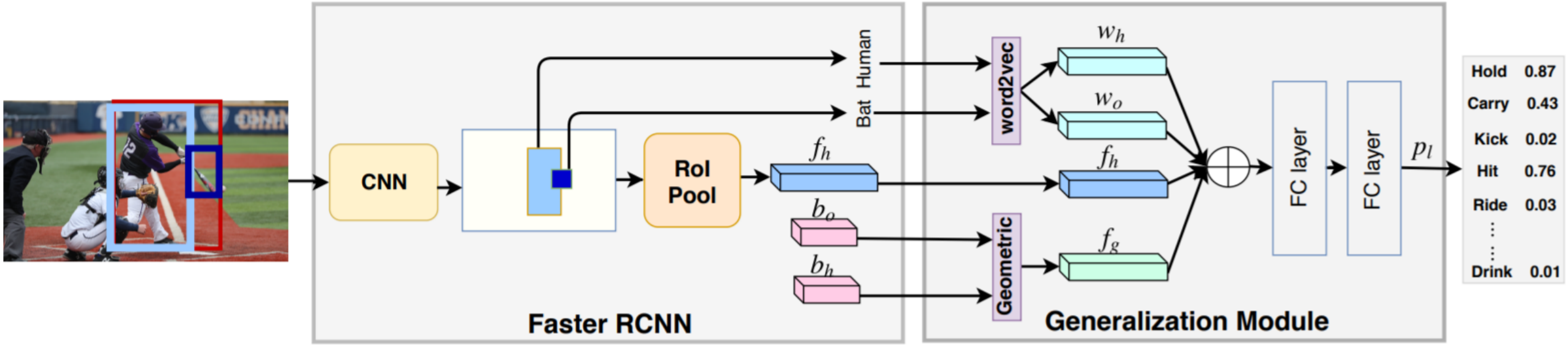

**Skip-gram : 중심 단어에서 주변 단어 예측**



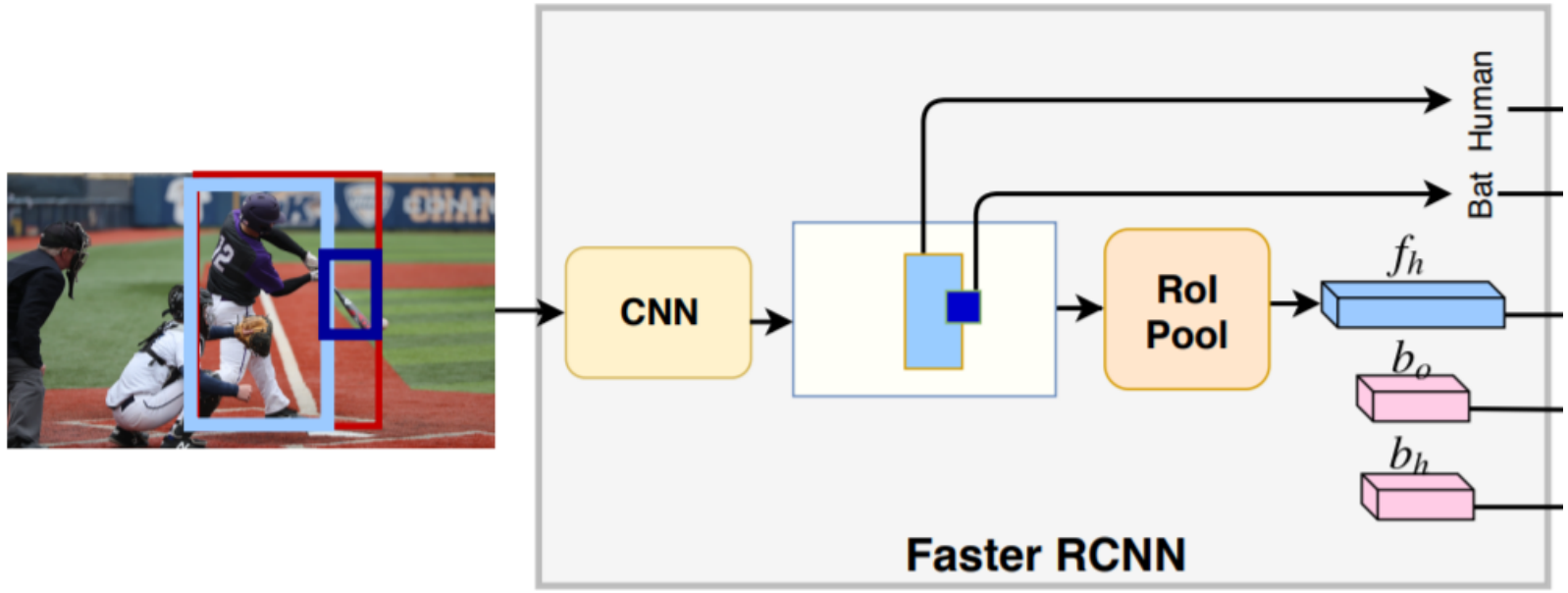

```
model_result2 = model.wv.most_similar("어벤져스")
print(model_result2)
```

```
[('스파이더맨', 0.8560965657234192), ('아이언맨', 0.8376990556716919), ('데어데블', 0.7797115445137024),
('인크레더블', 0.7791407108306885), ('스타트렉', 0.7752881050109863), ('엑스맨', 0.7738450765609741), ('슈
퍼맨', 0.7715340852737427), ('어벤저스', 0.7453964948654175), ('슈퍼히어로', 0.7452991008758545), ('다크나
이트', 0.7413955926895142)]
```

# 제안하는 모델 :

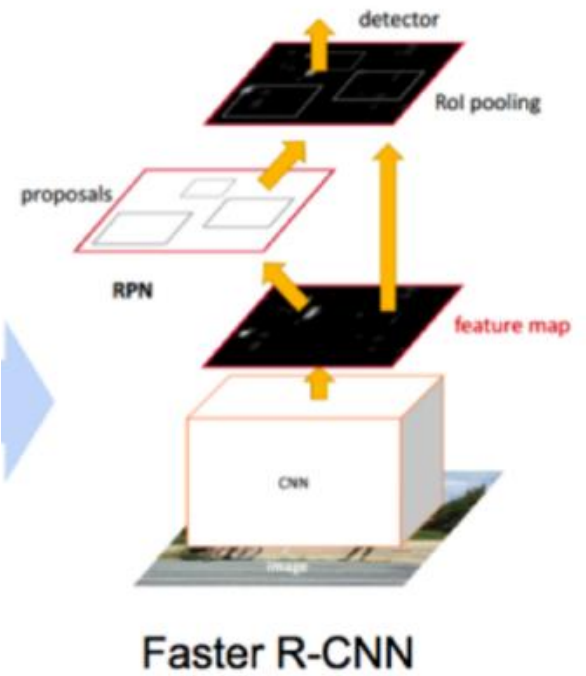

CNN
RoI Pool
Human
Bat
$f_h$
$b_o$
$b_h$
Faster RCNN
word2vec
Geometric
$w_h$
$w_o$
$f_h$
$f_g$
FC layer
FC layer
$p_l$
Hold 0.87
Carry 0.43
Kick 0.02
Hit 0.76
Ride 0.03
Drink 0.01
Generalization Module

# Object Detection :

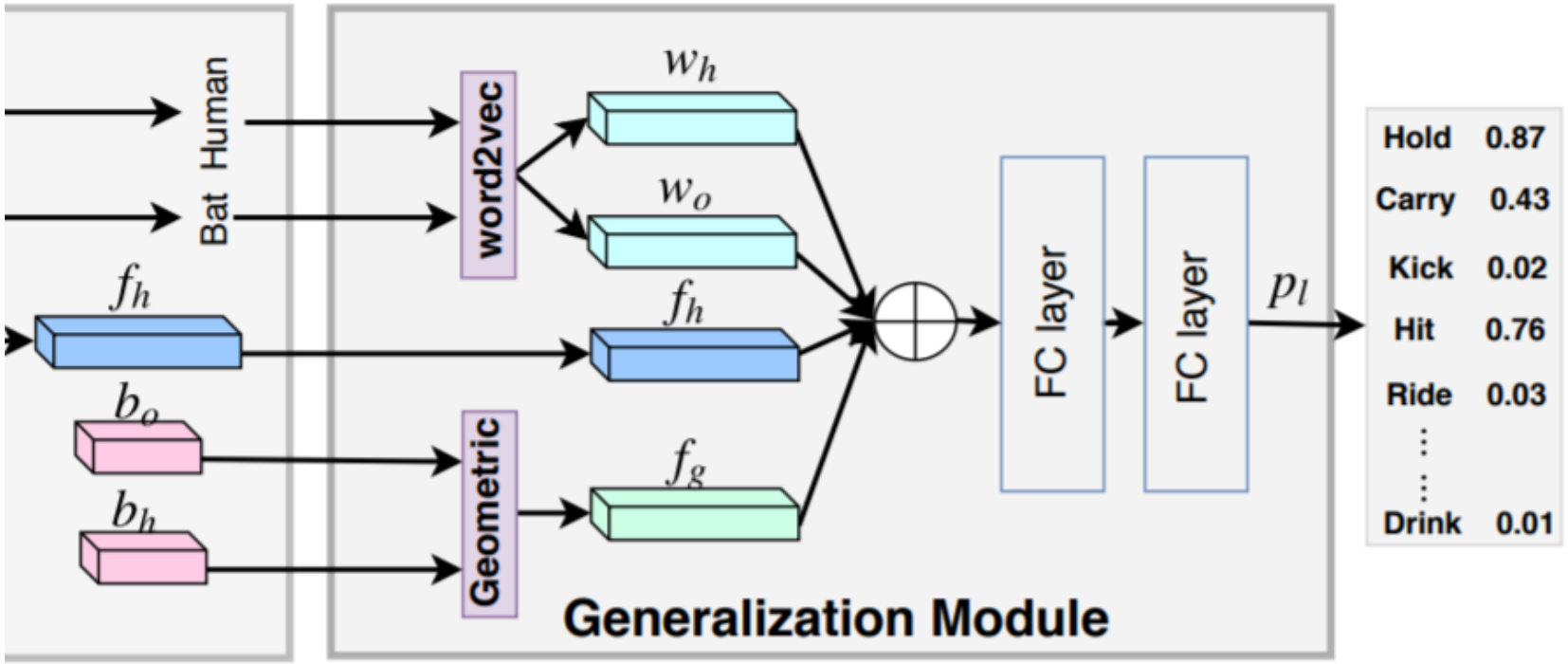



OID(Open Images dataset)로 트레이닝 된 Faster-RCNN

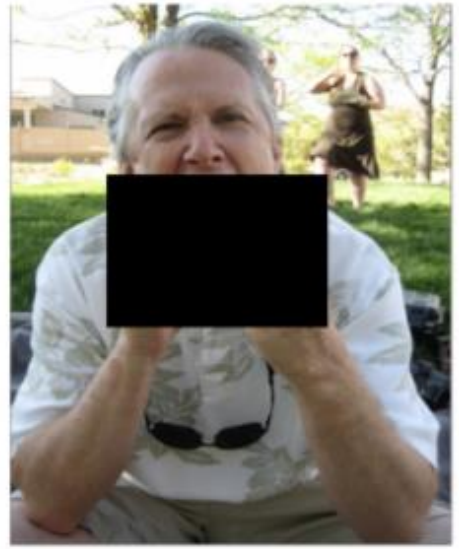
545 object categories



# Functional Generalization Module :




| | |
|---|---|
| Hold | 0.87 |
| Carry | 0.43 |
| Kick | 0.02 |
| Hit | 0.76 |
| Ride | 0.03 |
| ⋮ | |
| Drink | 0.01 |



⟨human, eat, __

Word2vec의 300-D를 사용하여 객체의 유사성을 이용하기 위한 인풋

$W_h$ : 남자, 여자, 소년, 소녀, 사람.... 등 사람을 임베딩 벡터로 세분화 사용

$W_o$ : 유사성 있는 여러 객체들의 임베딩 벡터를 사용

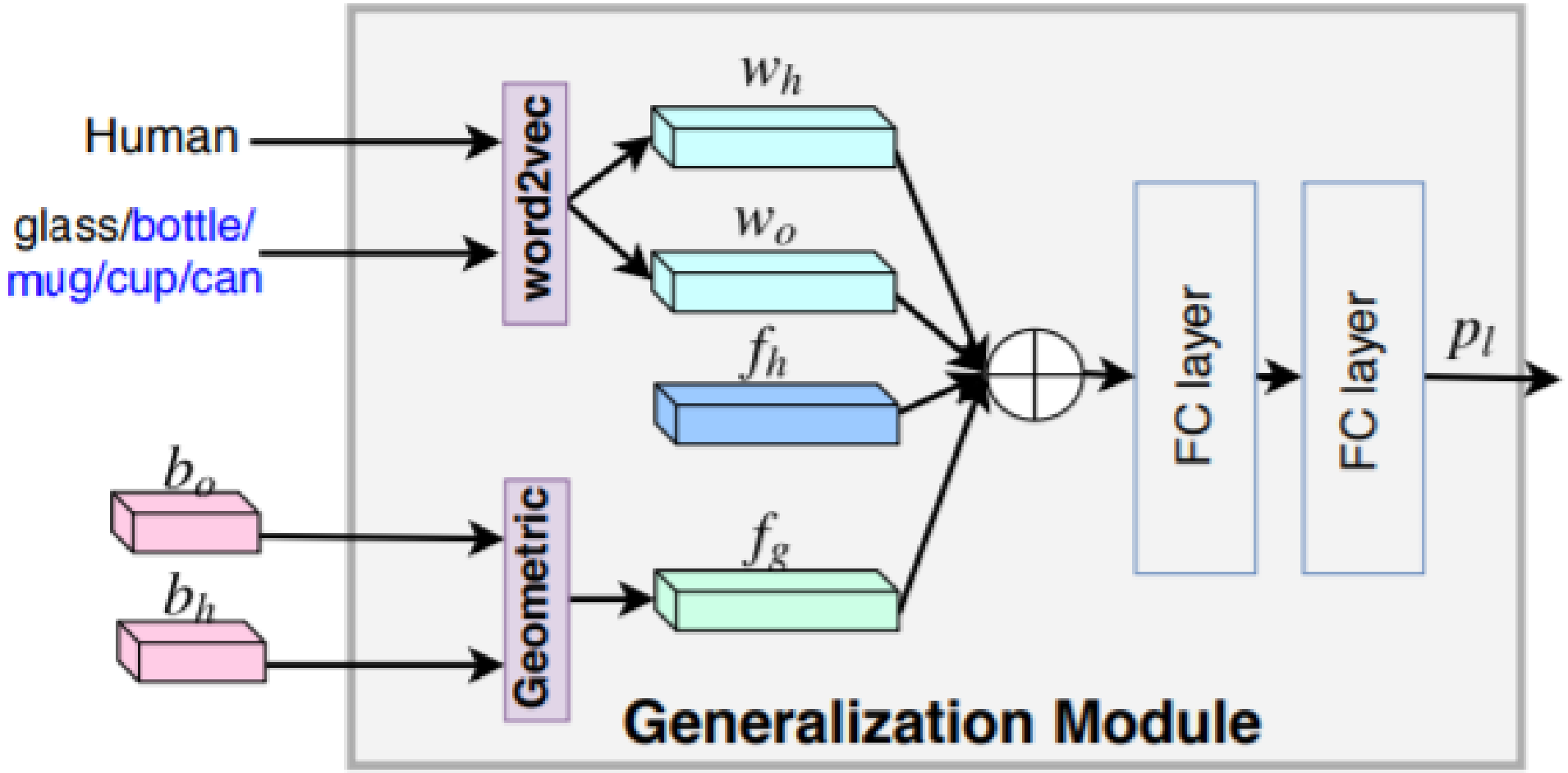
# 시각적, 기하학적 특징 인풋

$f_h$ : 인간-객체 쌍의 시각적 특징

$f_g$ : 인간, 객체의 바운딩 박스 좌표를 통해 기하학적 특징

$$f_g = \left[ \frac{x_1^h}{W}, \frac{y_1^h}{H}, \frac{x_2^h}{W}, \frac{y_2^h}{H}, \frac{A^h}{A^I}, \frac{x_1^o}{W}, \frac{y_1^o}{H}, \frac{x_2^o}{W}, \frac{y_2^o}{H}, \frac{A^o}{A^I}, \left( \frac{x_1^h - x_1^o}{x_2^o - x_1^o} \right), \left( \frac{y_1^h - y_1^o}{y_2^o - y_1^o} \right), \log \left( \frac{x_2^h - x_1^h}{x_2^o - x_1^o} \right), \log \left( \frac{y_2^h - y_1^h}{y_2^o - y_1^o} \right) \right] \quad (1)$$

# 일반화 모듈 훈련을 위해


Human
glass/bottle/
mug/cup/can
word2vec
$w_h$
$w_o$
$f_h$
$b_o$
$b_h$
Geometric
$f_g$
FC layer
FC layer
$p_l$
Generalization Module

# 데이터셋

HICO-DET
Images 47,776 (38,118, 9,658)
Objects 80 (airplane, apple…)
Verbs 117 (carry, catch…)
HOI 600 (airplane – board, direct, exit, fly…)

HOI Remark >= 150k

# 실험 결과 (HICO-Det)

| Method | Full (600 classes) | Rare (138 classes) | Non-Rare (462 classes) | # Params (millions) |
|---|---|---|---|---|
| Shen *et al.* (Shen et al. 2018) | 6.46 | 4.24 | 7.12 | - |
| HO-RCNN + IP (Chao et al. 2017) | 7.30 | 4.68 | 8.08 | - |
| HO-RCNN + IP + S (Chao et al. 2017) | 7.81 | 5.37 | 8.54 | - |
| InteractNet (Gkioxari et al. 2017) | 9.94 | 7.16 | 10.77 | - |
| iHOI (Xu et al. 2018) | 9.97 | 7.11 | 10.83 | - |
| GPNN (Qi et al. 2018) | 13.11 | 9.34 | 14.23 | - |
| ICAN (Gao, Zou, and Huang 2018) | 14.84 | 10.45 | 16.15 | 48.1 + 40.9 = 89.0 |
| Gupta *et al.* (Gupta, Schwing, and Hoiem 2019) | 17.18 | 12.17 | 18.68 | 9.2 + 63.7 = 72.9 |
| Interactiveness Prior (Li et al. 2019) | 17.22 | 13.51 | 18.32 | 35.0 + 29.0 = 64.0 |
| Peyre *et al.* (Peyre et al. 2019) | 19.40 | 15.40 | 20.75 | 21.8 + 40.9 = 62.7 |
| Ours | **21.96** | **16.43** | **23.62** | **3.1** + 48.0 = **51.1** |

# 실험 결과 (ZSD, HOI ZSD)

| Method | Unseen (120 classes) | Seen (480 classes) | All (600 classes) |
|---|---|---|---|
| Shen *et al.* (Shen et al. 2018) | 5.62 | - | 6.26 |
| Ours | **11.31**±1.03 | **12.74**±0.34 | **12.45**±0.16 |

| Method | Unseen (100 classes) | Seen (500 classes) | All (600 classes) |
|---|---|---|---|
| Ours | 11.22 | 14.36 | 13.84 |

# 실험 결과 (훈련할때 유사성 높은 객체 추가 수)

| r (Number of objects) | Full (600 classes) | Rare (138 classes) | Non-Rare (462 classes) |
|---|---|---|---|
| 0 | 12.72 | 7.57 | 14.26 |
| 3 | 13.70 | 7.98 | 15.41 |
| 5 | **14.35** | **9.84** | **15.69** |
| 7 | 13.51 | 7.07 | 15.44 |

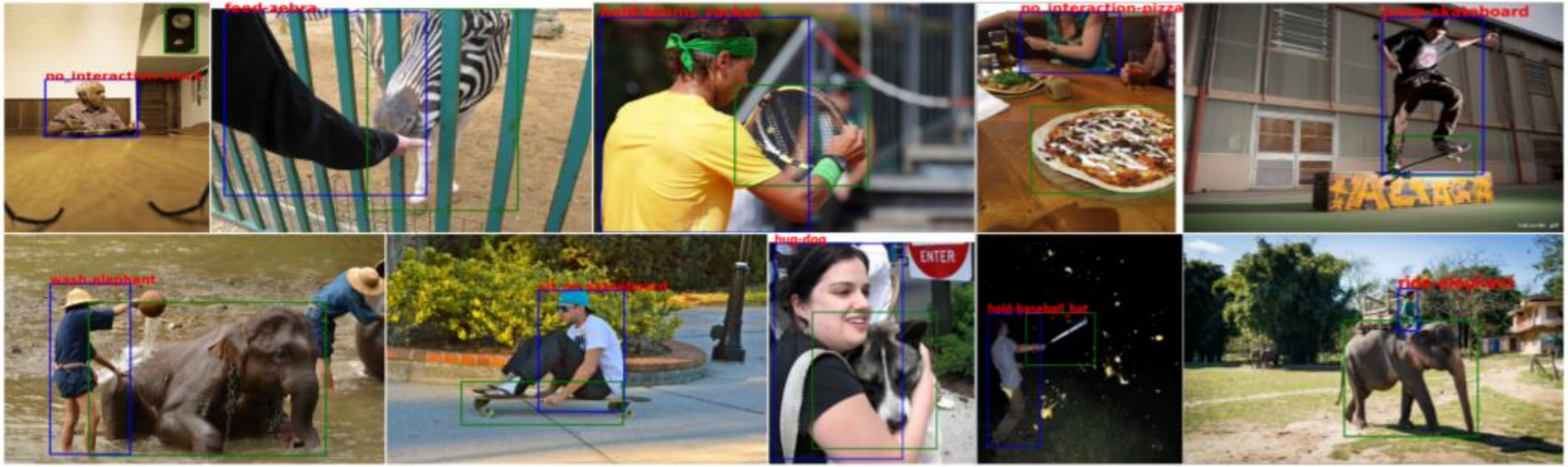
# 실험 결과 (클러스터링 알고리즘에 따른 실험)

| Clustering Algorithm | Full (600 classes) | Rare (138 classes) | Non-Rare (462 classes) |
|---|---|---|---|
| K means | **14.35** | **9.84** | 15.69 |
| Agglomerative | 14.05 | 7.59 | **15.98** |
| Affinity Propagation | 13.49 | 7.53 | 15.28 |

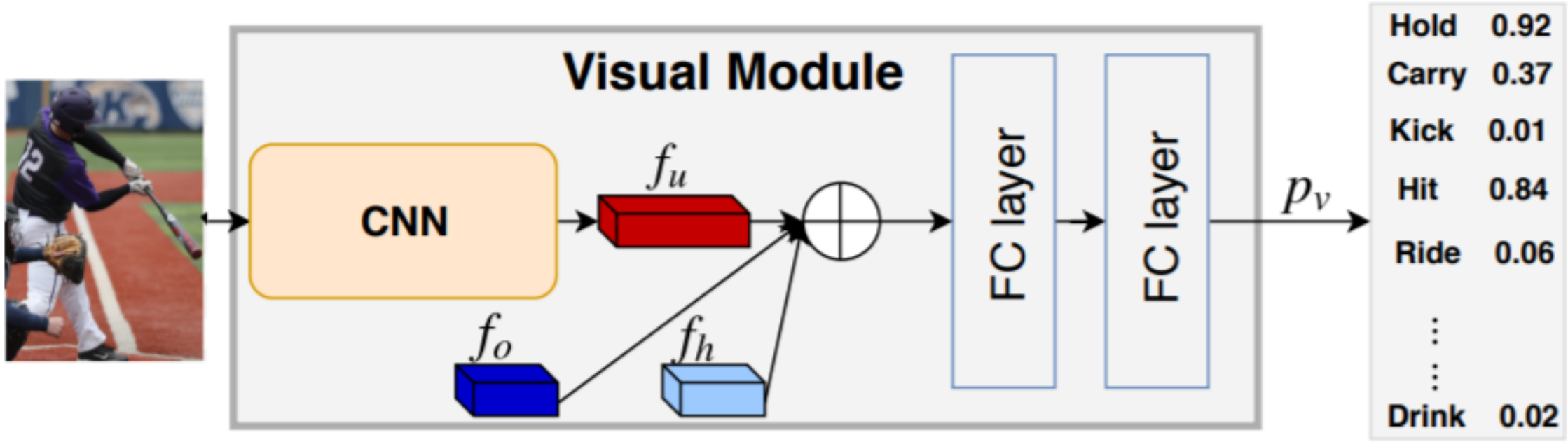
# 실험 결과 (인풋 절제 연구)

| Setting | Full (600 classes) | Rare (138 classes) | Non-Rare (462 classes) |
|---|---|---|---|
| Base | **14.35** | **9.84** | **15.69** |
| Base $-f_h$ | 12.15 | 4.87 | 14.33 |
| Base $-f_g$ | 12.43 | 8.02 | 13.75 |
| Base $-w_h - w_o$ | 12.23 | 5.23 | 14.32 |

# 실험 결과


feed-zebra
no_interaction-pizza
wash-elephant
hug-dog
ENTER
hold-baseball_bat
ride-elephant

# 추가 연구


Visual Module
CNN
$f_u$
$f_o$
$f_h$
FC layer
FC layer
$p_v$
Hold 0.92
Carry 0.37
Kick 0.01
Hit 0.84
Ride 0.06
Drink 0.02

# Q & A