https://arxiv.org/abs/2001.02302

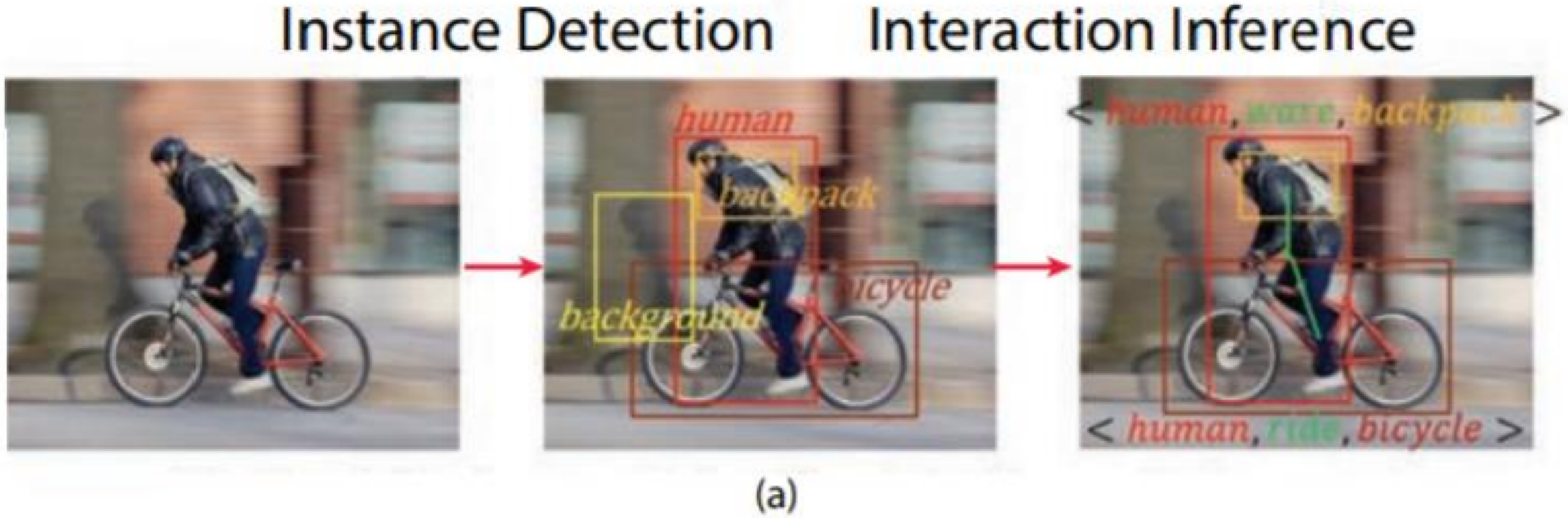
# Visual-Semantic Graph Attention Network for Human-Object Interaction Detection

Zhijun Liang, Yisheng Guan, and Juan Rojas
Guangdong University of Technonlogy
2020

인공지능 연구실
석사과정 구자봉

# 문제 정의 :
# HOI(Human Object Interaction)

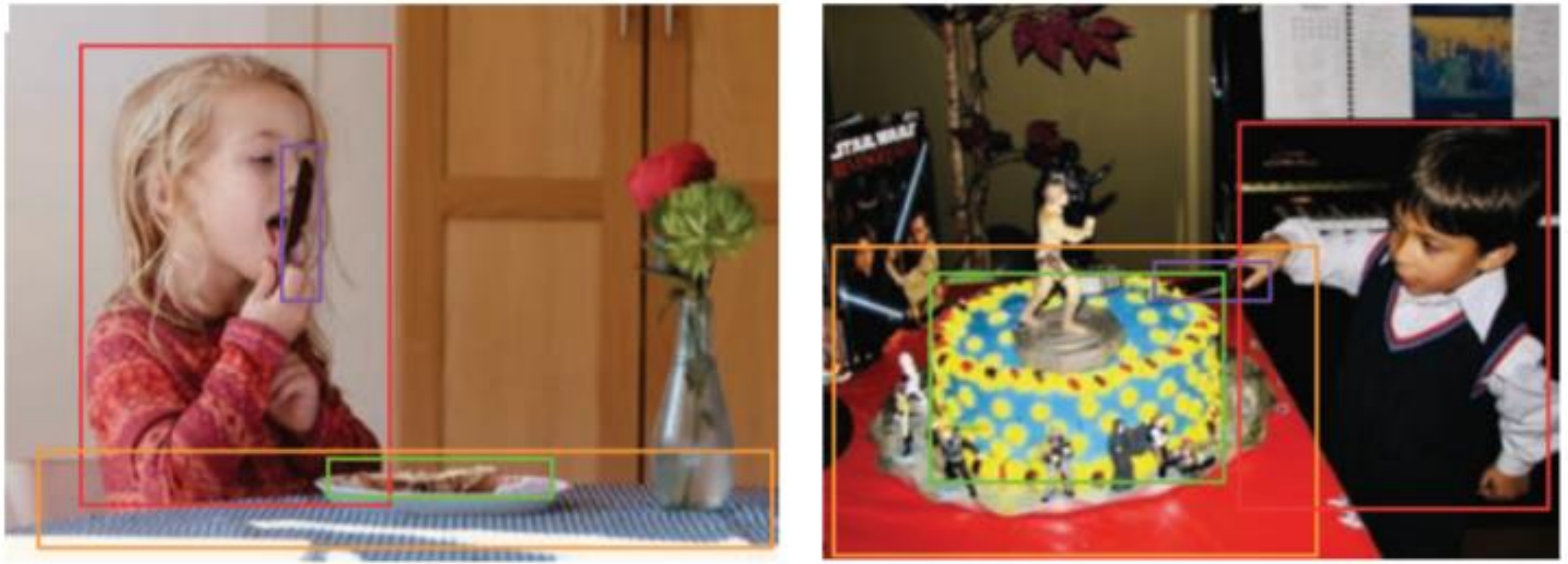
이미지에서 오브젝트 디텍션, 인간과 상호작용이 큰 객체쌍을 선택, 술어(상관관계)를 찾는 것이 목적




(a)

# 문제 제기 :

왼쪽은 명확히 <human,lick,knife>임,
오른쪽은 오브젝트 두개의 쌍으로는 조금 애매함 <human,hold,knife>, <human,?,cake>



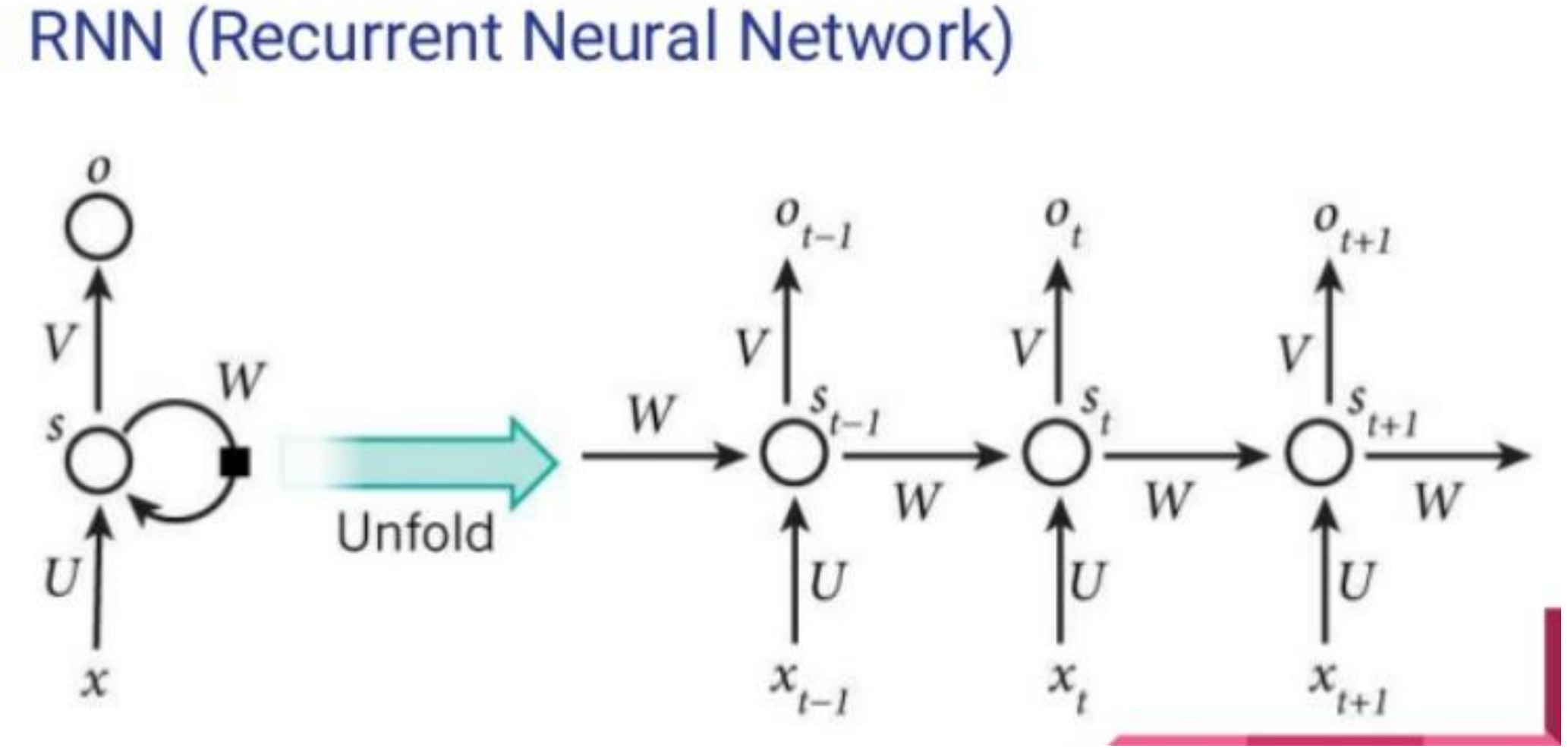
(b)

# 목표:

비주얼정보(이미지)와 시맨틱정보(의미론적)가 모두 포함된 그래프 기반 모델을 통해 HOI 문제 접근

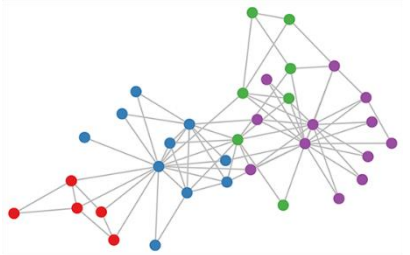
http://www.secmem.org/blog/2019/08/17/gnn/

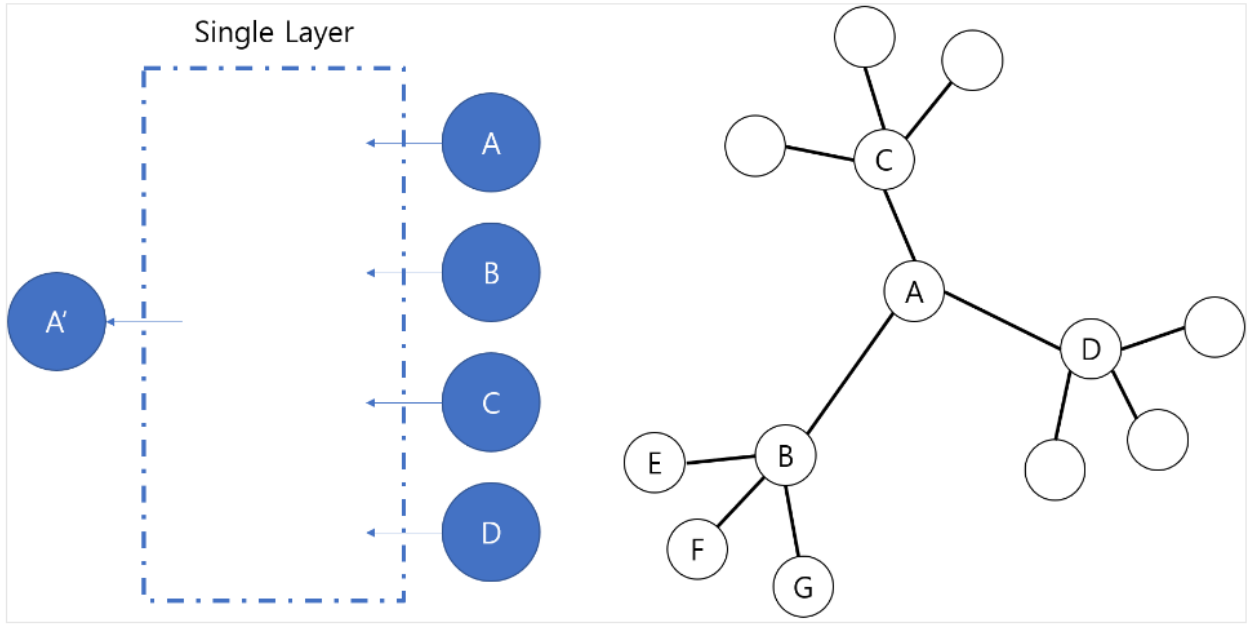
# Recurrent Neural Network(RNN):

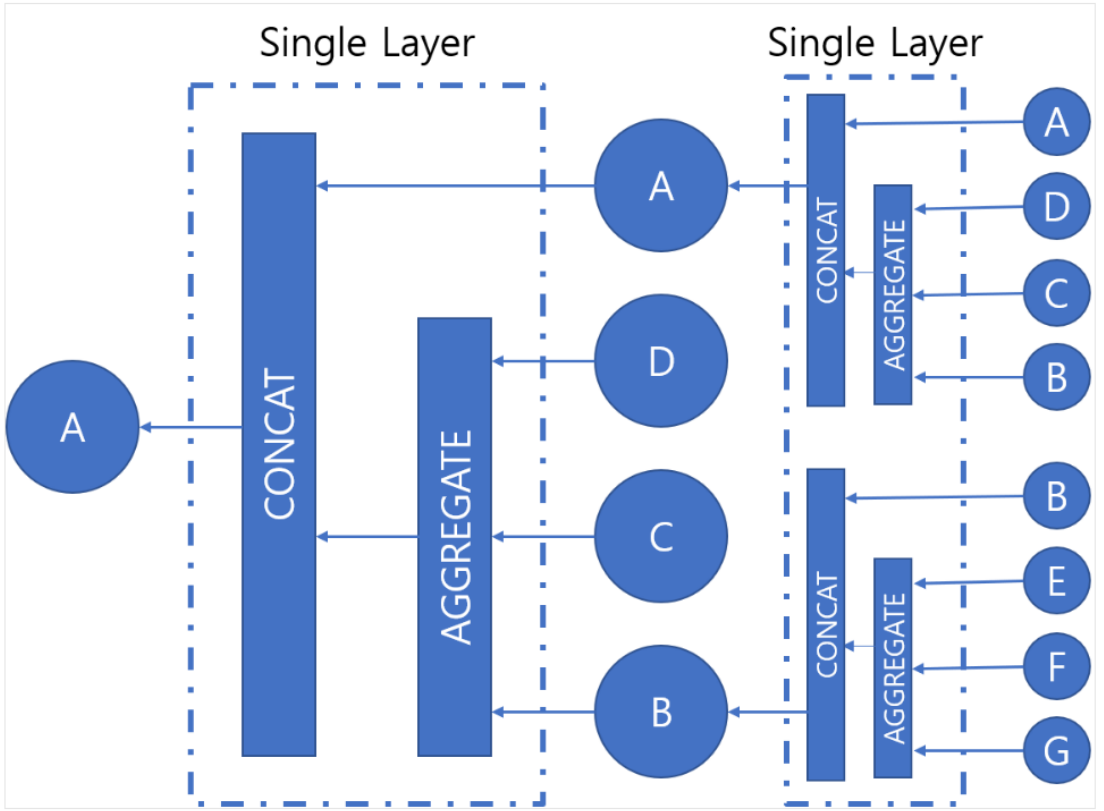


http://www.secmem.org/blog/2019/08/17/gnn/

# Graph Neural Network(GNN):





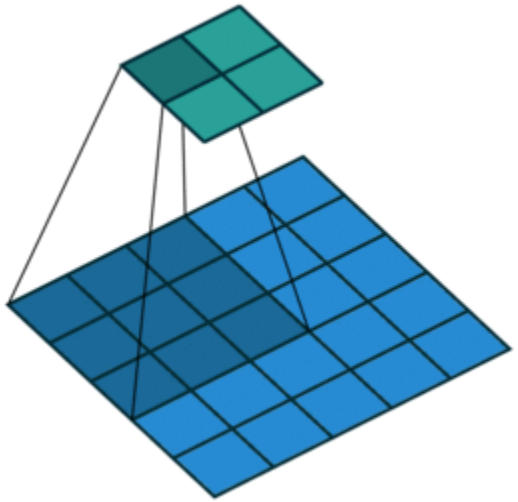




자기 자신을 주위의 노드들과 과거 자신의 feature 를 입력으로 새로운 자신을 임베딩하여 만듬

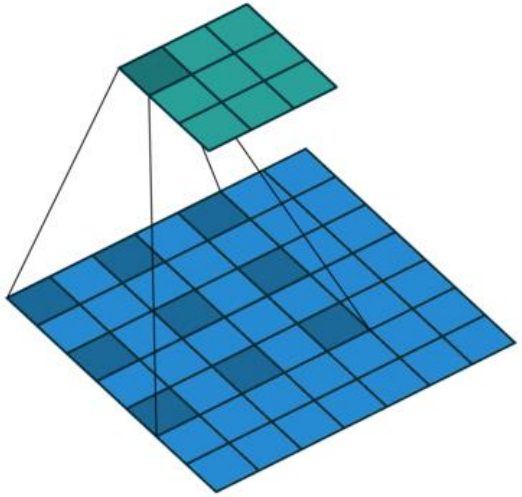
$$h_v^k = \sigma(W_k \sum_{\{u|\{u,v\}\in E\}} \frac{h_u^{k-1}}{|\{u|\{u,v\}\in E\}|} + B_k h_v^{k-1})$$

https://newsight.tistory.com/313

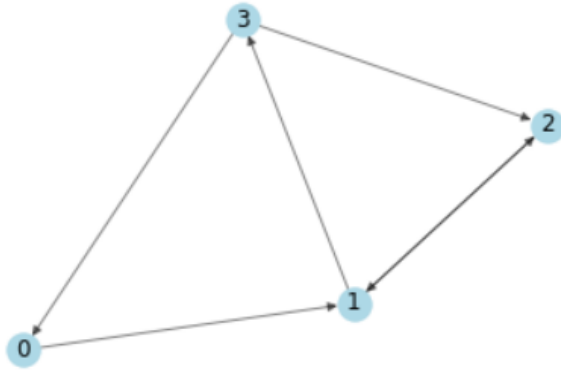
# Graph Convolutional Network(GCN):



CNN



CNN(dilation=2)




```
A = np.matrix([
    [0, 1, 0, 0],
    [0, 0, 1, 1],
    [0, 1, 0, 0],
    [1, 0, 1, 0]],
    dtype=float)
```

GCN
Adjacency Matrix

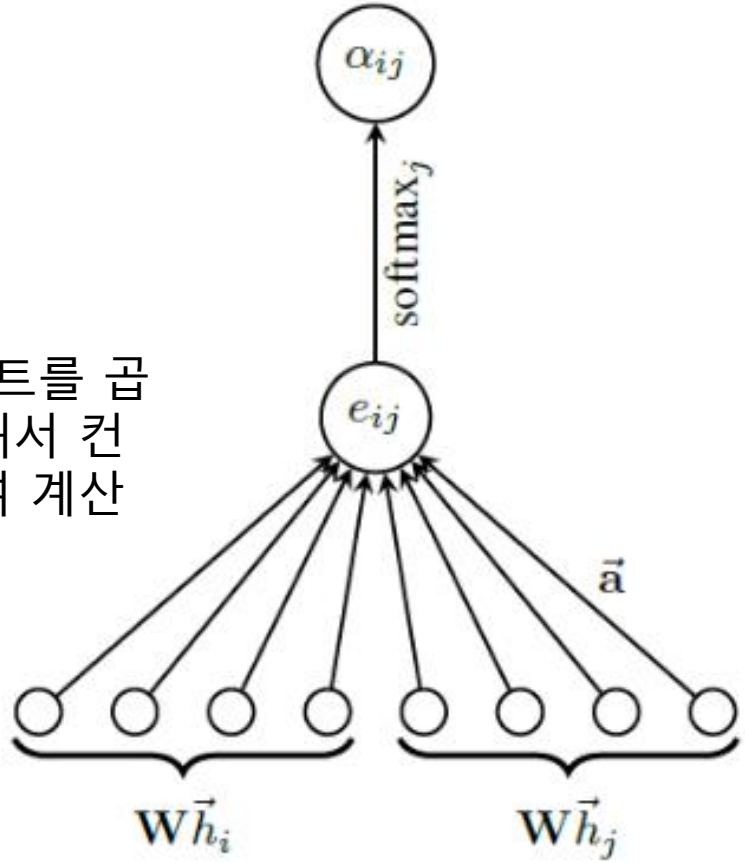
$$h_v^k = \sigma\left(W_k \sum_{\{u|\{u,v\}\in E\}\cup\{v\}} \frac{h_u^{k-1}}{\sqrt{|\{w|\{v,w\}\in E\}||\{w|\{u,w\}\in E\}|}}\right)$$

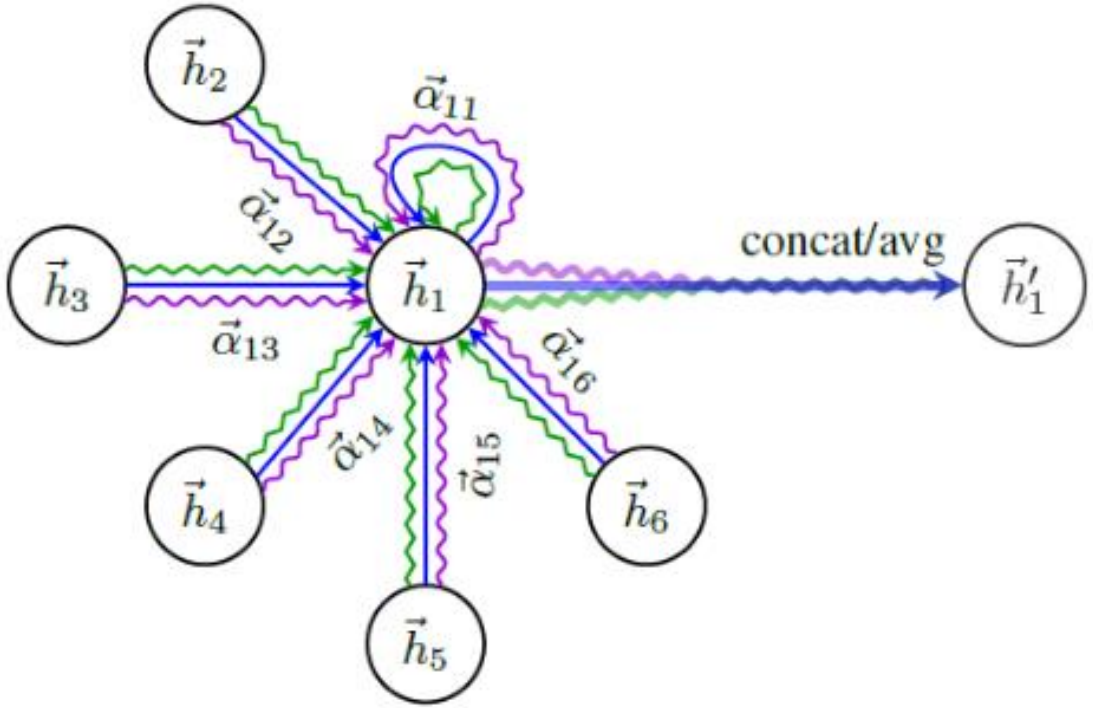
https://arxiv.org/pdf/1710.10903.pdf

https://newsight.tistory.com/313

# Graph Attention Networks(GATs):

i피쳐와 j피쳐에 웨이트를 곱하고 리니어 임베딩해서 컨켓한 피쳐와 내적하여 계산

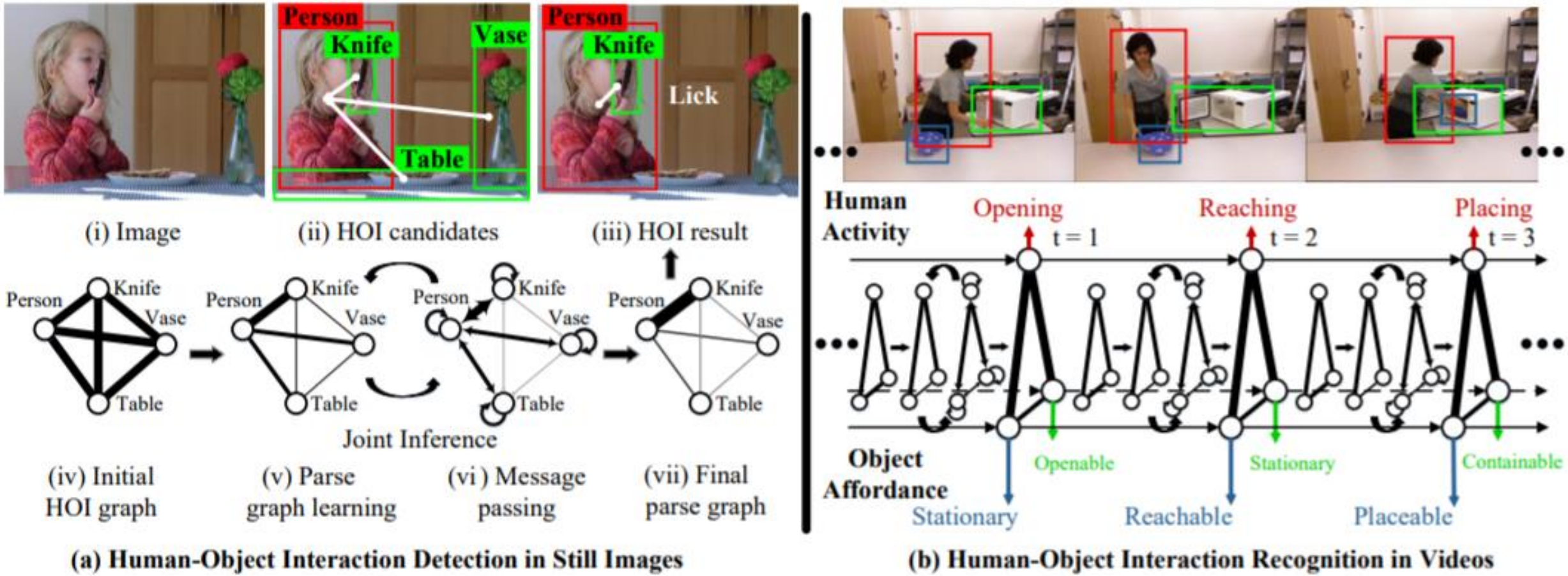


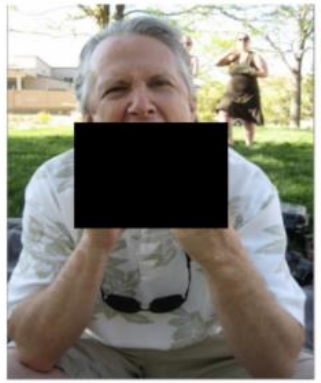

$$\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\mathbf{W}^k\vec{h}_j\right)$$

https://arxiv.org/pdf/1710.10903.pdf

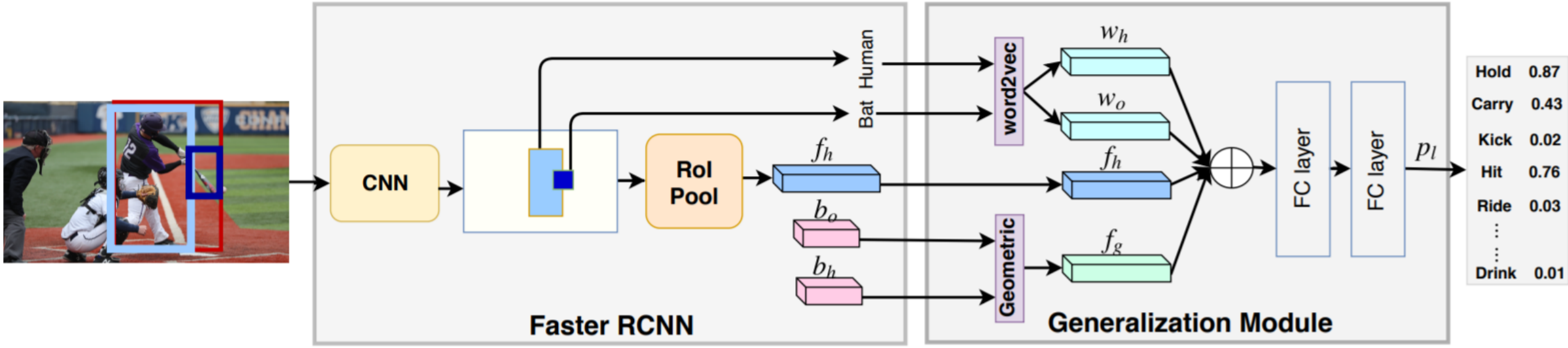
# Learning Human-Object Interactions by Graph Parsing Neural Networks (GPNN):



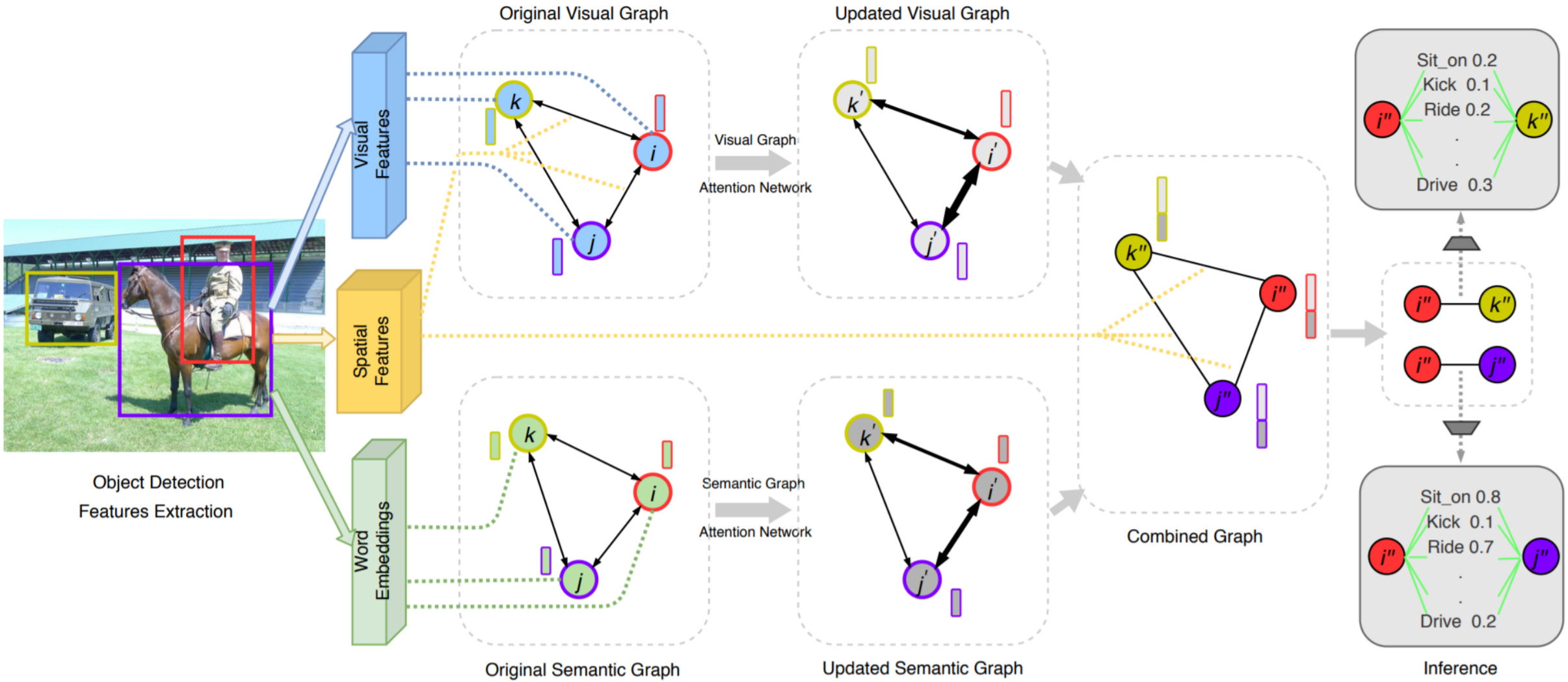
# Detecting Human-Object Interactions via Functional Generalization



⟨human, eat, __



# Visual-Semantic Graph Attention Network for Human-Object Interaction Detection(VS-GATs)



# VS-GATs 목표 정의

$$G = (V, E)$$
$$e_{i,j} = (v_i, v_j) \in E$$

그래프 정의

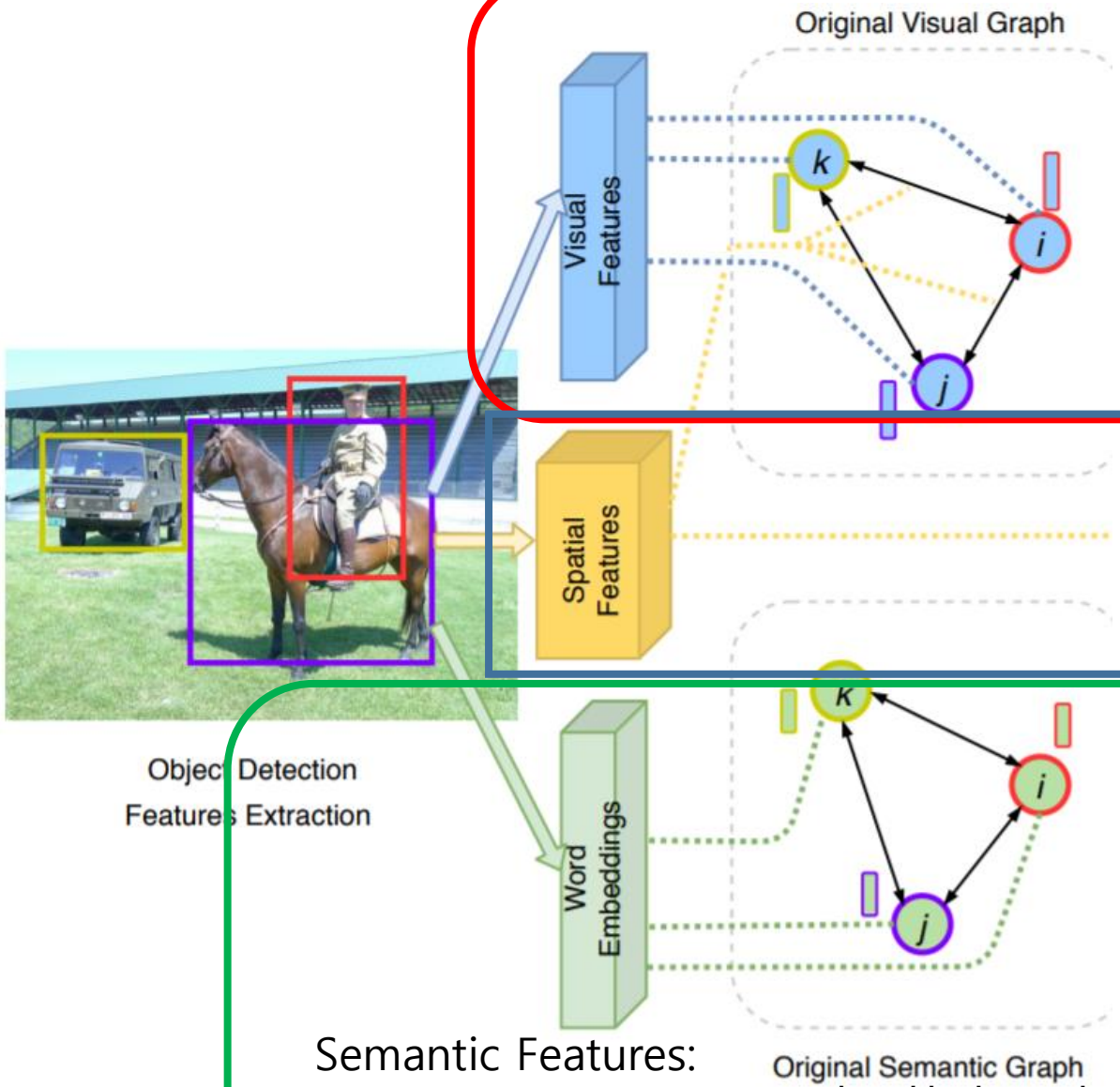
$$\text{image } I \qquad \text{class labels } \mathbf{R}$$

이미지와 라벨 주어짐

$$P(\mathbf{R}, G_C, G_V, G_S \mid I) = P(G_V, G_S \mid I)$$
$$P(G_C \mid G_V, G_S, I)$$
$$P(R \mid G_C, G_V, G_S, I).$$

이미지와 라벨 주어졌을 때 Visual Graph와 Semantic Graph를 만들고, 둘을 컨켓하여 Combine Graph 생성, 이를 통해 R 유추

# VS-GATs 피쳐




Visual Features:
FasterRCNN (RestNet-50-FPN)에서 오브젝트 추출
ROI pooling layer의 Feature를 사용함

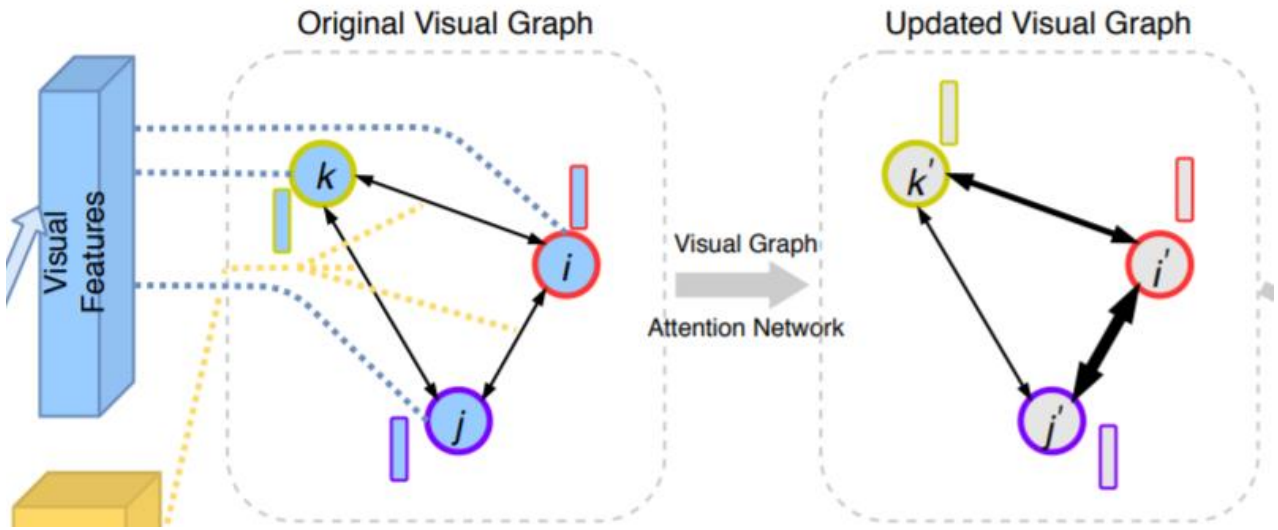
Spatial Features:

$$s_{rs} = \left[\frac{x_i}{W}, \frac{y_i}{H}, \frac{x_j}{W}, \frac{y_j}{H}, \frac{A}{A^I}\right]$$

$$s_{rp} = \left[\left(\frac{x_i - x_i^{'}}{x_j^{'} - x_i^{'}}\right), \left(\frac{y_i - y_i^{'}}{y_j^{'} - y_i^{'}}\right), \log\left(\frac{x_j - x_i}{x_j^{'} - x_i^{'}}\right), \log\left(\frac{y_j - y_i}{y_j^{'} - y_i^{'}}\right), \frac{x_c - x_c^{'}}{W}, \frac{y_c - y_c^{'}}{H}\right]$$

Semantic Features:
word2vec vectors 300D 사용하여 유사도에 따른 Feature 사용

# VS-GATs의 Visual GATs 수식




FasterRCNN, ROI 풀링 피쳐로 초기 노드피쳐와 에찌피쳐를 사용하여 그래프 생성

$$\mathbf{h}_{e_{ij}} = f_{edge}([\mathbf{h}_{v_i}, \mathbf{s}_{ij}, \mathbf{h}_{v_j}])$$

스페셜 피쳐와 노드피쳐를 사용해에찌함수를 사용해 히든피쳐를 만듬

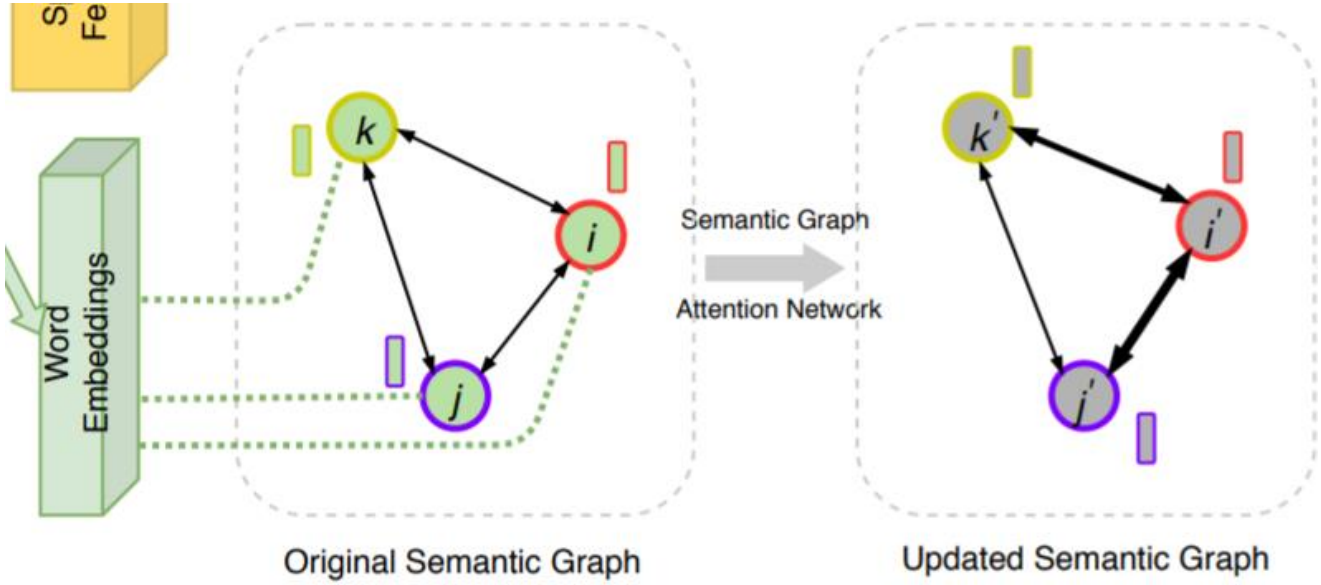
$$\mathbf{z}_{h_i} = \sum_{j \in \mathcal{N}_i} \alpha_{ij}(\mathbf{h}_{v_j} \oplus \mathbf{h}_{e_{ij}})$$

주위의 모든 노드와 히든피쳐를 계산한 값들을 다 더함

$$\tilde{\mathbf{h}}_{v_i} = f_{update}([\mathbf{h}_{v_i}, \mathbf{z}_{h_i}])$$

자신을 업데이트 함

# VS-GATs의 Semantic GATs 수식




Word2vec 300D를 통해 유사도 피쳐를 사용하여 그래프 생성

$$\alpha'_{ij} = softmax(f'_{attn}(f'_{edge}([\mathbf{w}_i, \mathbf{w}_j]))).$$

비쥬얼 그래프에서 사용된 에찌 함수와 어텐션 함수로 에찌의 가중치 연산

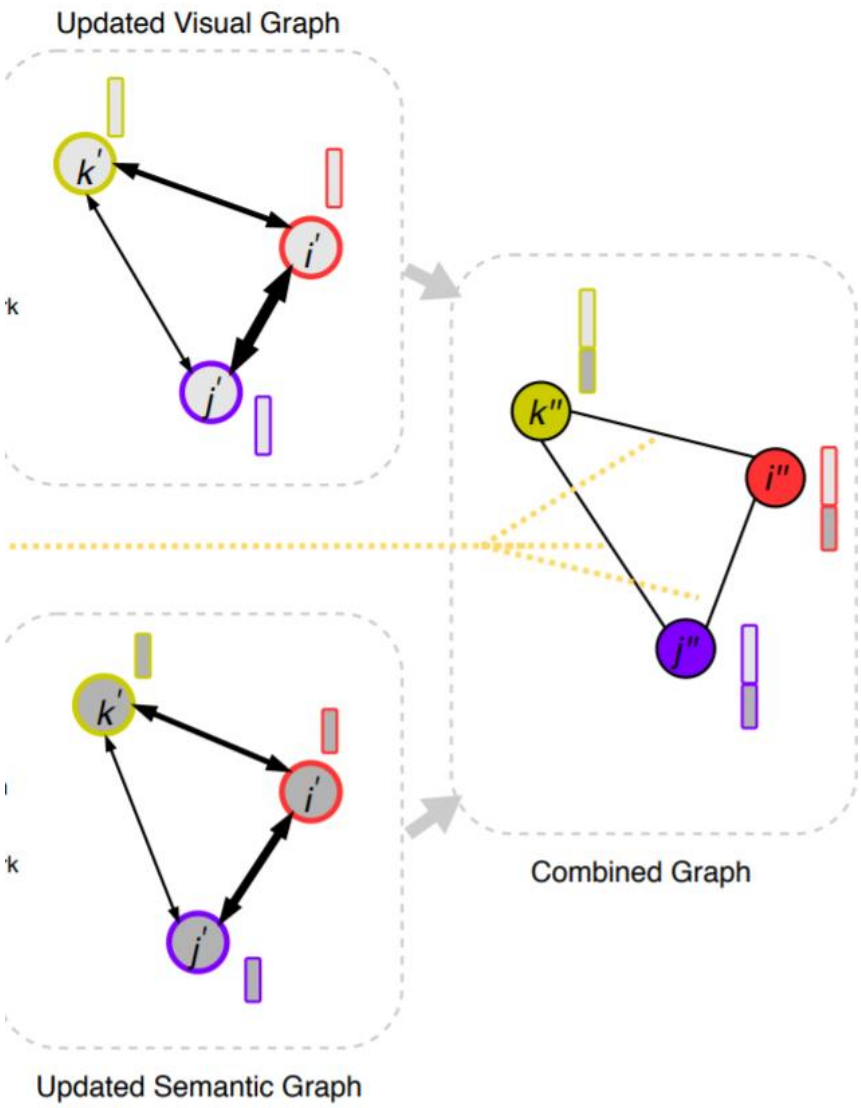
$$\mathbf{z}_{w_i} = \sum_{j \in N_i} \alpha'_{ij} \mathbf{w}_j.$$

주위의 모든 노드와 히든피쳐를 계산한 값들을 다 더함

$$\tilde{\mathbf{w}}_i = f'_{update}([\mathbf{w}_i, \mathbf{z}_{w_i}]).$$

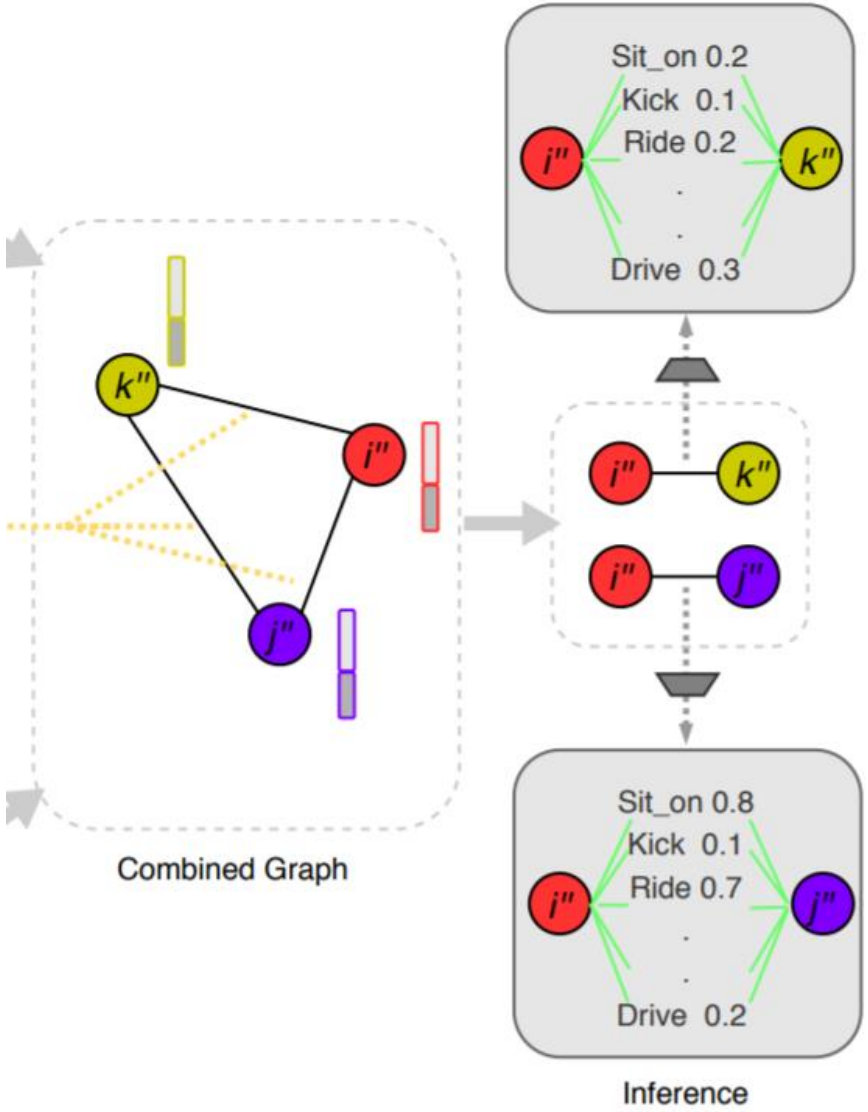
자신을 업데이트 함

# VS-GATs의 Combined GATs




컨켓

# VS-GATs의 Readout and Inference 수식




가장 강하게 연결된 사람과 객체 쌍이 선택 되면 다중 분류 문제임

$$\mathbf{S}_a = sigmoid(f_{readout}(a))$$

$$\mathbf{S_R} = s_h * s_o * \mathbf{S_a}.$$

# VS-GATs의 Training

$$\mathcal{L} = \sum_i \sum_j BCE(\mathbf{S}_{a_{ji}}, Y_{ji}^{label})$$

멀티 클래스 크로스 엔트로피 사용

# Dataset

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

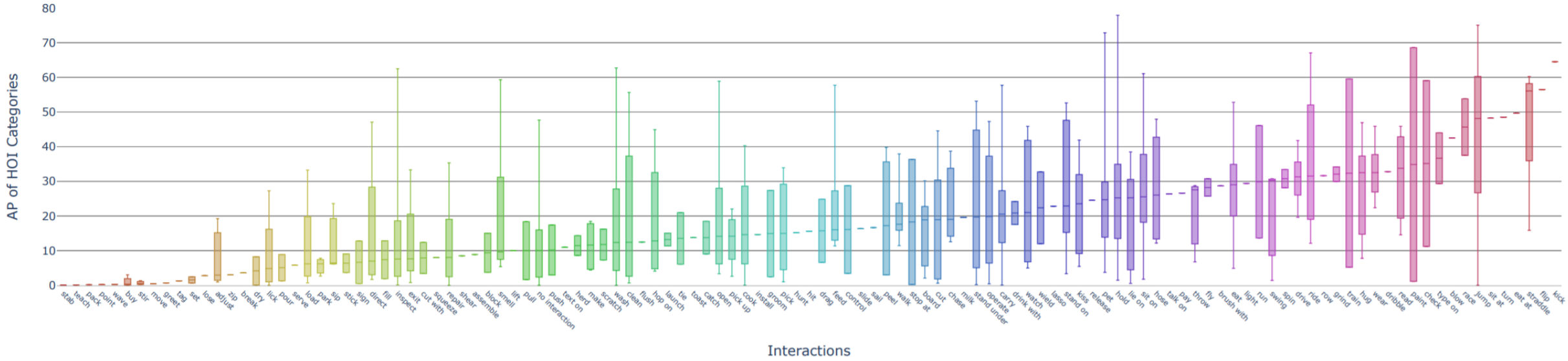
HOI Remark >= 150k

# 실험 HICO-DET

| Method | Object Detector | Full(600)↑ | Rare(138)↑ | Non-Rare(462)↑ |
|---|---|---|---|---|
| InteractNet [9] | Faster R-CNN with ResNet-50-FPN | 9.94 | 7.16 | 10.77 |
| GPNN [28] | Deformable ConvNets [7] | 13.11 | 9.34 | 14.23 |
| iCAN [8] | Faster R-CNN with ResNet-50-FPN | 14.84 | 10.45 | 16.15 |
| Xu *et al.* [36] | Faster R-CNN with ResNet-50-FPN | 14.70 | 13.26 | 15.13 |
| Gupta *et al.* [12] | Faster R-CNN with ResNet-152 | 17.18 | 12.17 | 18.68 |
| Li et al. $RP_{T2}C_D$ [18] | Faster R-CNN with ResNet-50-FPN | 17.22 | 13.51 | 18.32 |
| PMFNet [34] | Faster R-CNN with ResNet-50-FPN | 17.46 | 15.65 | 18.00 |
| Peyre *et al.* [26] | Faster R-CNN with ResNet-50-FPN | 19.40 | 14.60 | **20.90** |
| **Ours(VS-GATs)** | Faster R-CNN with ResNet-50-FPN | **19.66** | **15.79** | 20.81 |

Table 1. mAP performance comparison with SOTA on the HICO-DET test set.

# 실험



# 실험 with out 절제실험

| Method | Full↑ | Rare↑ | Non-Rare↑ |
|---|---|---|---|
| **Ours(VS-GATs)** | **19.66** | **15.79** | **20.81** |
| 01 $G_V$ only | 18.81 | 13.96 | 20.26 |
| 02 $G_S$ only | 14.61 | 11.76 | 15.46 |
| 03 w/o attention | 19.01 | 14.12 | 20.47 |
| 04 w/o spatial features in $G_C$ | 18.52 | 14.28 | 19.78 |
| 05 Message passing in $G_C$ | 19.23 | 14.31 | 20.70 |
| 06 Unified V-S graph | 19.39 | 14.84 | 20.75 |

Table 2. mAP performance for various ablation studies.

# 결론

주요 사람 – 오브젝트의 인터렉션뿐만 아니라 subsidiary relations 도 활용한 연구를 진행함

# Q & A