Supplementary Material for "Hierarchical Graph Attention Network for Visual Relationship Detection"

Li Mi, Zhenzhong Chen*
School of Remote Sensing and Information Engineering, Wuhan University, China
{milirs,zzchen}@whu.edu.cn

1. Dataset

The examples of VRD and VG dataset are presented in Figure 1.

1.1. Visual Relationship Detection (VRD)

There are 70 predicate categories in VRD dataset.

Verb: attach to, carry, contain, cover, drive, drive on, eat, face, feed, fly, follow, hit, hold, kick, lean on, look, lying on, park behind, park next, park on, play with, pull, rest on, ride, sit behind, sit next to, sit on, sit under, skate on, sleep next to, sleep on, stand behind, stand next to, stand on, stand under, talk, touch, use, walk, walk beside, walk next to, walk past, walk to, watch, wear.

Spatial: above, adjacent to, behind, below, beneath, beside, in, in the front of, inside, near, next to, on, on the left of, on the right of, on the top of, outside of, over, under.

Preposition: across, against, at, by, has, with.

Comparative: taller than.

1.2. Visual Genome (VG)

There are 100 predicate categories in VG dataset.

Verb: adorn, attach to, belong to, build into, carry, cast, catch, connect to, contain, cover, cover in, cover with, cross, cut, drive on, eat, face, fill with, fly, fly in, grow in, grow on, hang in, hang on, hit, hold, hold by, lay in, lay on, lean on, look at, mount on, paint on, park, play, print on, pull, read, reflect in, rest on, ride, say, show, sit at, sit in, sit on, stand behind, stand on, standing by, standing in, standing next to, support, surround, swing, throw, touch, use, walk, walk in, walk on, watch, wear, wear by, write on.

Spatial: above, behind, below, beneath, between, in, in front of, in middle of, inside, near, next to, on, on back of, on bottom of, on side of, on top of, outside, over, under, underneath.

Preposition: across, against, along, around, at, beside, by, for, from, have, of, part of, to, with.

Comparative: small than, tall than.

*Corresponding author: Zhenzhong Chen.

Figure 1. Visualization of relationship annotation in VRD and VG dataset. For each triplet, the orange bounding box denotes the region of subject, while the yellow bounding box denotes the region of object. The predicate is represented as a green line, which connects the subject and the object.

2. Graph Attention Visualization

The visualization of graph attention is shown in Figure 2 and Figure 3 for object-level graph attention and triplet-level graph attention, respectively.

Figure 2. Visualization of object-level graph attention maps. The red bounding box is the reference region and the other four bounding boxes shown in each image are the top-4 attended regions.

2.1. Object-level Graph Attention

Figure 2 demonstrates the effectiveness of object-level graph attention. As is shown in Figure 2, object attention weight varies with different reference object. For example, comparing the first and second image of the last row, the object "man" is more close to the object "jacket" and "pants", while the object "mountain" is more likely to interact with the object "snow" and "house".

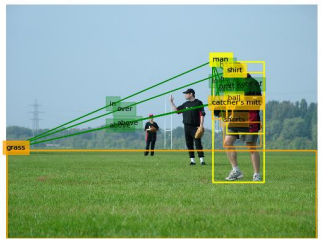
2.2. Triplet-level Graph Attention

From Figure 3 we can see, interactions among triplets are significant context information for VRD. For example, the attention maps of "dog-frisbee" and "water-bank" are different, which indicates the other triplets in the image will have different effects on these two triplets.

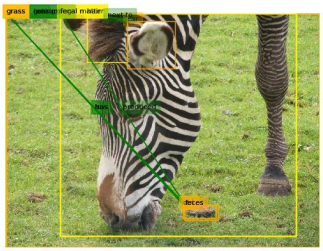
3. More Examples

More results of HGAT are shown in Figure 4. Green, yellow and red color denotes the correct triples, correct but unannotated triples and failed triples, respectively. In some examples, there are redundant edges and missing edges in the initial graph. The qualitative results demonstrate that prior knowledge and attention mechanism alleviate the detrimental effects of inaccurate graph.

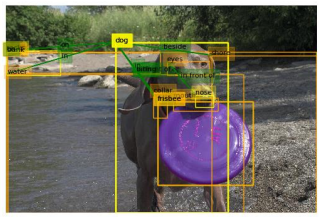
Figure 3. Visualization of triplet-level graph attention maps. The red bounding box and the blue bounding box represents the subject region and object region, respectively.



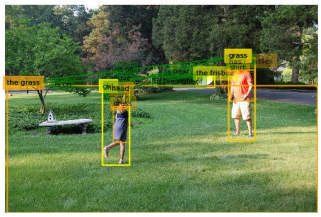
man has shirt
man wears shorts
man in grass
dog has nose
shirt near grass



grass around zebra
zebra eating grass
zebra has mouse
zebra has leg
zebra has ear

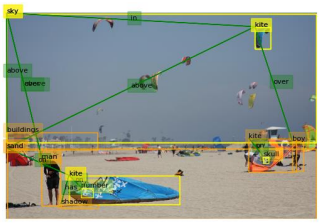


frisbee in front of dog
dog on sand
dog near water
dog has nose
dog near rocks

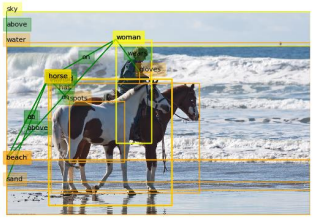


man on grass
man has shirt
frisbee next to man
girl wears dress
girl has frisbee

beach has sand
sand on beach
waves next to beach
water near beach
dog has nose
man near water

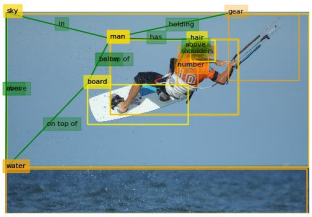


kite in sky
frisbee next to man
kite over man
buildings near sand
man has shirt



sky above water
woman on horse
woman on beach
horse near horse
horse in water

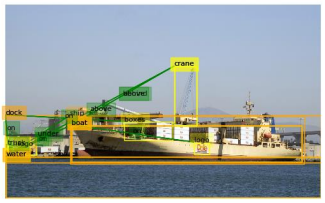
person with kite
person on grass
tree next to tree
man in grass
trees near grass



sky above water
board below man
man on top of board
number on shirt
horse in water

car on road
sky above road
car on street
light on road
car has lamp

airplane has tail
airplane has engine
airplane has wing
airplane on ground
airplane near mountain



boat on water
sky above water
water under boat
boxes on boat
boat near boat

Figure 4. More predicate prediction results of HGAT. Green, yellow and red color denotes the correct triples, correct but unannotated triples and failed triples, respectively.