Attention-Driven Dynamic Graph Convolutional Network for Multi-Label Image Recognition

Jin Ye[1*], Junjun He[1,2*], Xiaojiang Peng[1⋆], Wenhao Wu[1], and Yu Qiao[1†]

[1] ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
[2] School of Biomedical Engineering, the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China.

In supplementary material, we visualize more examples to illustrate whether SAM can locate semantic targets and what relations dynamic graph has learned for a single image.

For each example, (a) is the input image. (b) is the dynamic matrix $\mathbf{A}_d$ of (a). Specifically, the value of $\mathbf{A}_d^{c1;c2}$ donates the relation of $c1$ and $c2$ when $c1$ appears. And we can find $\mathbf{A}_d^{c1;c2}$ is not equal to $\mathbf{A}_d^{c2;c1}$ easily. (c) is category-specific activation maps of (a). The caption of activation map (e.g. Car: 1.00) means that the final classification score of the category "car" is "1.00" .

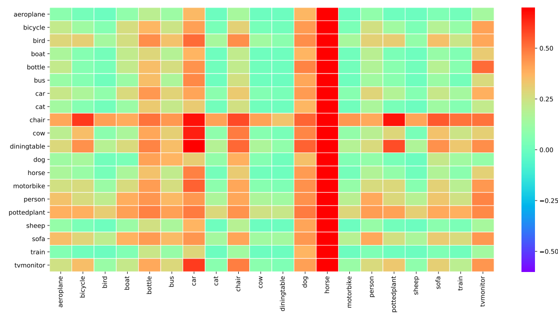
ADD-GCN learns a dynamic graph for each images. And we can observe that the labels of each image have strong relation values in the dynamic graph even though they have lower co-occurrence possibilities in the real world. For example, the probability that "dog" and "bottle" come together is very low in the real world or in an common image. But we can find that the relevant scores of "dog" and "bottle" ($\mathrm{A}_d^{dog;bottle}$ and $\mathrm{A}_d^{bottle;dog}$) rank top in each row ($\mathbf{A}_d^{dog}$ and $\mathbf{A}_d^{bottle}$) from Fig 1(b). The scores indicate that they have strong relation in Fig 1(a). Similar results can be found in other examples.

Table 1: The dictionary of dynamic matrix on MS-COCO. Each cell is a map of index to category of dynamic matrix on MS-COCO.

0 airplane	1 apple	2 backpack	3 banana	4 baseball bat	5 baseball glove	6 bear	7 bed	8 bench	9 bicycle
10 bird	11 boat	12 book	13 bottle	14 bowl	15 broccoli	16 bus	17 cake	18 car	19 carrot
20 cat	21 cell phone	22 chair	23 clock	24 couch	25 cow	26 cup	27 dining table	28 dog	29 donut
30 elephant	31 fire hydrant	32 fork	33 frisbee	34 giraffe	35 hair drier	36 handbag	37 horse	38 hot dog	39 keyboard
40 kite	41 knife	42 laptop	43 microwave	44 motorcycle	45 mouse	46 orange	47 oven	48 parking meter	49 person
50 pizza	51 potted plant	52 refrigerator	53 remote	54 sandwich	55 scissors	56 sheep	57 sink	58 skateboard	59 skis
60 snowboard	61 spoon	62 sports ball	63 stop sign	64 suitcase	65 surfboard	66 teddy bear	67 tennis racket	68 tie	69 toaster
70 toilet	71 toothbrush	72 traffic light	73 train	74 truck	75 tv	76 umbrella	77 vase	78 wine glass	79 zebra

⋆ Equally-contributed first authors. †Corresponding author (yu.qiao@siat.ac.cn)

(a) Input image

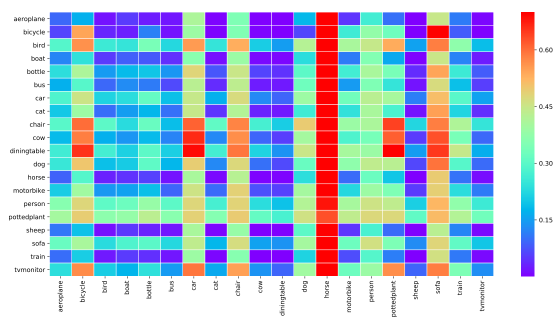


(b) Dynamic matrix

(c) Category-specific activation maps

Fig. 1: Example on VOC2007. Labels are “bottle” and “dog”.

(a) Input image

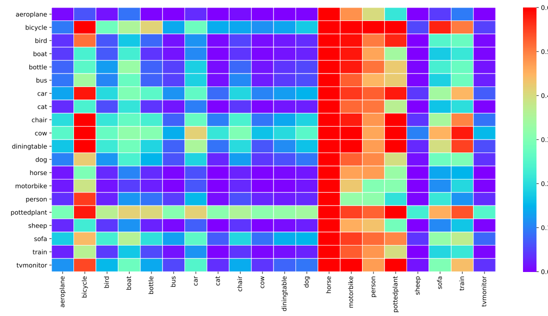


(b) Dynamic matrix

(c) Category-specific activation maps

Fig. 2: Example on VOC2007. Labels are “bicycle” and “chair”.

(a) Input image

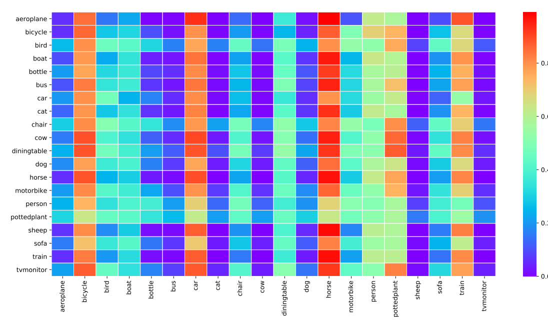


(b) Dynamic matrix

(c) Category-specific activation maps

Fig. 3: Example on VOC2007. Labels are “person”, “motorbike” and “bicycle”.

(a) Input image

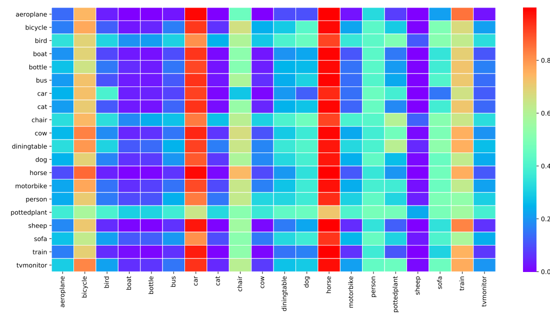


(b) Dynamic matrix

(c) Category-specific activation maps

Fig. 4: Example on VOC2007. Labels are “car”, “bicycle” and “person”.

(a) Input image

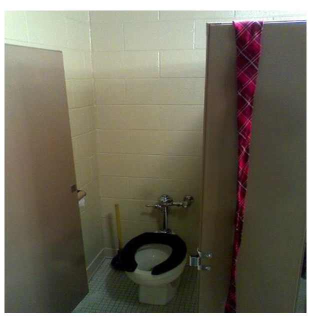


(b) Dynamic matrix

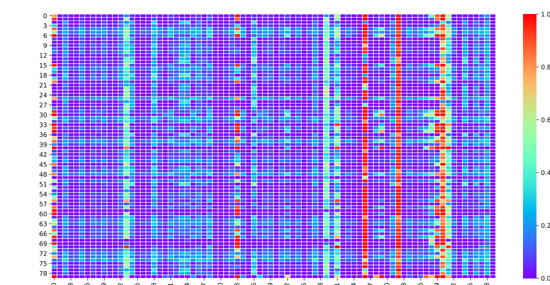
Car: 1.00 Horse: 0.99 Person: 0.99 Chair: 0.01

(c) Category-specific activation maps

Fig. 5: Example on VOC2007. Labels are “car”, “horse” and “person”.



(a) Input image



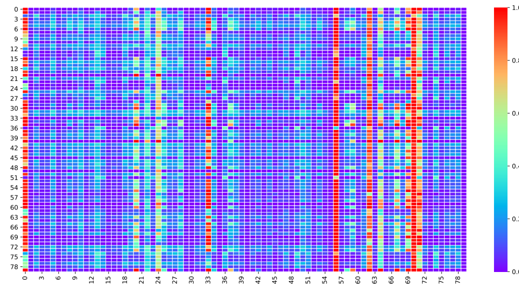
(b) Dynamic matrix

Tie: 1.00 Toilet: 1.00 Sports ball: 0.00 Toothbrush: 0.00

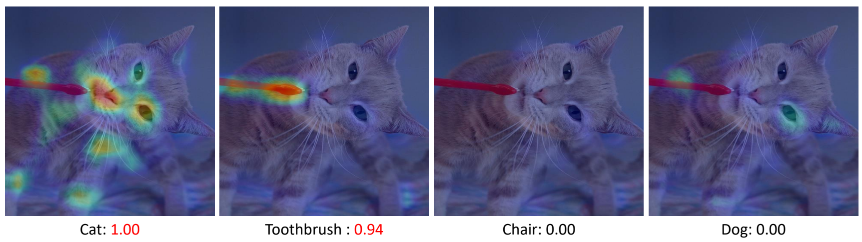
(c) Category-specific activation maps

Fig. 6: Example on MS-COCO. Labels are “tie” and “toilet”.

(a) Input image

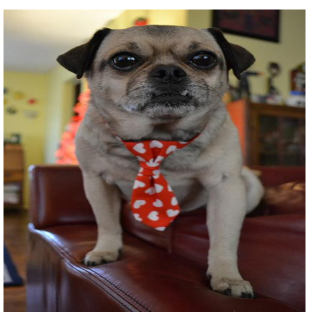


(b) Dynamic matrix

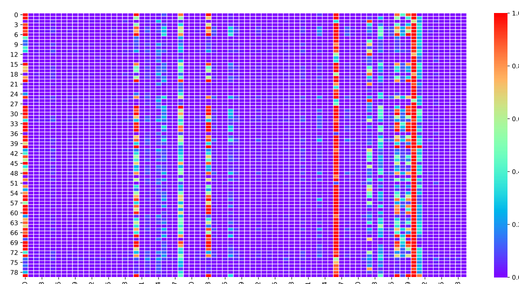


(c) Category-specific activation maps

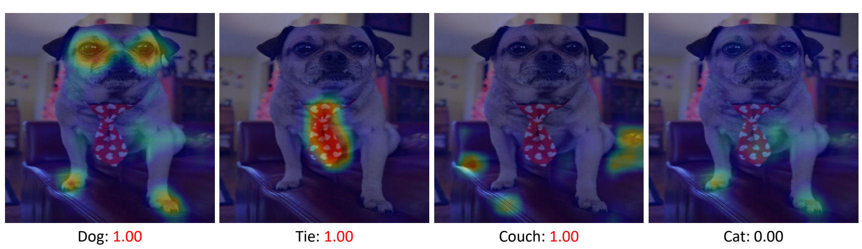
Fig. 7: Example on MS-COCO. Labels are “cat” and “toothbrush”.



(a) Input image



(b) Dynamic matrix



(c) Category-specific activation maps

Fig. 8: Example on MS-COCO. Labels are “dog”, “tie” and “couch”.