Divide and Conquer: Answering Questions with Object Factorization and Compositional Reasoning (Supplementary Materials)

Shi Chen Qi Zhao
Department of Computer Science and Engineering,
University of Minnesota
{chen4595, qzhao}@umn.edu

Our supplementary materials consist of additional experimental results that demonstrate the effectiveness of the proposed method:

1. We perform experiments to validate the generalizability of our method in two additional scenarios, *i.e.,* questions about small or rare objects, respectively. (Section 1)

2. We present an ablation study on the model design of our prototypical neural module network. (Section 2)

3. We provide more detailed results on how our proposed prototypes characterize objects based on their semantic relationships (Section 3).

1. Additional Generalizability Experiments

One of the primary focuses of our study is to develop computational methods that can generalize toward various real-world scenarios. In the main paper, we show that, with object factorization and compositional reasoning, our approach is able to generalize to questions with novel concepts or variable data distributions. To further demonstrate its advantages, in this section we carry out experiments on two additional generalization settings: (1) questions with small objects, and (2) questions with rare objects. The former one considers ten novel objects with small sizes in the VQA dataset, *e.g.,* bottle, bird, tie. While the latter reorganizes the GQA training and evaluation data by removing training questions with the top 30% least frequent objects and only considering evaluation questions with the removed objects. Results show that our method is advantageous in both scenarios, *i.e.,* POEM: 52.11 and 55.78 vs XNM: 50.77 and 54.62.

2. Ablation Study on Model Design

The proposed prototypical neural module (POEM) network incorporates discriminative prototypes within different neural modules (*i.e., Find*, *Filter*, *Describe*), and considers prototypes attended over time with a semantic memory module. In this section, we investigate the effectiveness of different components in the model by experimenting with two of its variants: (1) A variant that only utilizes prototypes in an independent *Prototype* neural module (*i.e.,* POEM-Ind), where *Prototype* is an additional module alongside the existing ones in the XNM [3] baseline, and (2) Our model without the semantic memory module (POEM w/o Mem). Experimental results in Table 1 show that both variants lead to inferior performance, demonstrating the advantages of the design of our full POEM model.

Table 1. Comparative results of our full model POEM and its variants. Best results are highlighted in bold.

	VQA [1]		GQA [2]		Novel-VQA [4]
	Known	Novel	Known	Novel	Novel-Q
POEM-Ind	63.51	54.25	55.01	52.99	56.66
POEM w/o Mem	62.49	53.89	58.39	58.60	59.41
POEM	**63.80**	**54.82**	**60.60**	**59.71**	**60.73**

3. Additional Results on Object Characterization

In our main paper, we show that the proposed object factorization method is capable of learning semantically plausible prototypes. These prototypes encode abundant information about the characteristics of different objects, and can be used to cluster them into different groups with distinct semantic relationships (Table 4 of the main paper). This section provides the more detailed results of the clustering experiment, which contain the full list of objects divided into 30 groups. The results in Table 2 and Table 3 further demonstrate the usefulness of our prototypes for capturing the key characteristics of objects and correlating the relevant ones.

Table 2. Different groups of objects that are clustered based on their similarity with prototypes.

Group	Objects
1	cup, saucer, glass, beer, mug, juice, beverage, liquid, smoothie, coffee, syrup, coffee cup, wine, milk, wineglass, soda, yogurt, tea, milkshake, ...
2	fork, knives, spoon, knife, silverware, utensil, ladle, chopstick, tongs, spatula, butter knife, scissors, whisk, sword, earbuds
3	spectator, umpire, batter, catcher, crowd, net, player, baseball, match, stadium, bleachers, dugout, team, soccer ball
4	bed, pillowcase, sofa, pillow, bedspread, bedroom, headboard, comforter, couch, sheet, ottoman, armchair, mattress, bedding
5	water, sand, swimming pool, sea, rock, ocean, boulders, lake, beach, shore, river, cliff, ice, dock, stone, puddle, pond, mud, seaweed, canoe, harbor
6	vegetable, bowl, lettuce, salad, cabbage, pepper, broccoli, carrot, soup, spinach, cereal, rice, dinner, asparagus, beans, pasta, basil, celery, garnish, ...
7	pole, building, church, clock, roof, tower, barn, clock tower, house, statue, bridge, cone, post, antenna, cross, flag, streetlight, tunnel, city, bricks, balcony, ...
8	outlet, faucet, toilet, bathroom, towel, sink, shower, seat, lid, toilet paper, dispenser, tissue, tiles, bathtub, drain, dryer, urinal, soap dish, paper towel, ...
9	pants, shoe, sneakers, jeans, sandal, glove, snowboard, sock, snowboarder, boot, ski, skateboarder, skateboard, snow pants, leggings, skater, vacuum, heel, shoelace, knee pad
10	collar, hat, cap, scarf, backpack, t-shirt, shorts, glasses, pajamas, sweater, jersey, uniform, sweatshirt, bracelet, dress, skirt, apron, robe, bandana, undershirt, cloths, ...
11	bottle, can, vase, container, jar, candle, candle holder, canister, water bottle, alcohol, saltshaker, spray can, dish soap, hand soap, bottle cap, shampoo, wineglasses, ...
12	cake, crumbs, ice cream, muffin, coin, cupcake, candies, dessert, cream, donuts, sprinkles, frosting, sponge, pastries, chocolate, sugar, balloon, icing, cookie, sour cream, sushi, ...
13	table, chandelier, fireplace, window, ceiling, rug, curtain, chair, coffee table, floor, door, fan, carpet, frame, side table, cabinet, mirror, entrance, shelf, dresser, mat, heater, ...
14	cauliflower, onion, beef, meat, mushroom, egg, potatoes, chicken, tofu, pineapple, nut, potato, zucchini, shrimp, sausage, pickles, peanut, coconut, corn, mozzarella, ...
15	snow, ground, grass, field, fence, hill, meadow, zoo, dirt, forest, mountain, yard, family, hillside, park, hay, plain, mound, farm, pasture, lawn, desert, fog
16	cheese, bacon, napkin, dish, pizza, fish, food, butter, toast, tablecloth, sauce, sandwich, topping, bread, ham, tray, gravy, biscuit, platter, breakfast, hamburger, hot dog, bun, ...
17	tail, kite, bird, beak, seagull, sail, swan, duck, goose, feathers, parrot, parachute, dolphin, pigeon, flamingo, eagle, shark, peacock, geese
18	man, boy, girl, suit, lady, person, clothes, coat, men, child, snowsuit, gentleman, people, passenger, guy, surfer, tourist, pedestrian, mother, athlete, ...
19	tomato, strawberry, pepperoni, olive, berry, grape, blueberry, raspberry, cherry, raisin, blackberry, beet, eggplant, meatballs
20	airplane, propeller, airport, cockpit, aircraft, jet, helicopter, terminal, shuttle, whale
21	cat, figurine, doll, dog, panda, panda bear, bear, toy, nose, teddy bear, eyes, monkey, lips, kitten, baby, rubber duck, polar bear, fur, pig, ...
22	sticker, newspaper, picture, paper, sign, number, book, tape, drawing, cd, paint, package, letter, arrow, label, magazine, logo, display, painting, poster, word, ...
23	orange, grapefruit, fruit, banana, watermelon, kiwi, apple, lemon, lime, cucumber, produce, papaya, squash, pear, pumpkin, melon, avocado, mango, pomegranate, ...
24	motorcycle, bike, tire, bicycle, train, car, minivan, cart, vehicle, van, truck, taxi, train car, locomotive, SUV, firetruck, tractor, scooter, gas station, school bus, ...

Table 3. Different groups of objects that are clustered based on their similarity with prototypes (cont'd).

Group	Objects
25	pipe, pen, wire, cord, stick, toothbrush, chain, rope, brush, cable, straw, microphone, hose, guitar, pencil, bat, tool, comb, instrument, gun, toothbrushes, steering wheel, ...
26	elephant, giraffe, cow, calf, animal, zebra, sheep, donkey, lion, lamb, goat, bull, herd, bison, wool, deer, antelope, ostrich, rhino, moose
27	bush, flower, trunk, plant, leaf, branch, horn, tree, bushes, branches, leaves, pine tree, feeder, palm tree, garden, rose, mane, daisy, bouquet, garland, ...
28	box, drawer, microwave, counter, dishwasher, oven, stove, kitchen, bucket, sack, bag, trashcan, toaster, burner, appliance, suitcase, refrigerator, pot, teapot, basket, ...
29	binder, desk, screen, keyboard, laptop, television, phone, notebook, controller, computer, monitor, headphones, mouse pad, printer, speaker, camera, game, ...
30	street, path, platform, train station, walkway, pavement, road, gravel, sidewalk, runway, railroad, parking lot, crosswalk, skate park, station, highway, intersection, barrier, ...

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, pages 2425–2433, 2015. 1

[2] Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6693–6702, 2019. 1

[3] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8368–8376, 2019. 1

[4] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *CVPR*, pages 5628–5637, 2021. 1