ViCo: Word Embeddings from Visual Co-occurrences
(Supplementary Material)

Tanmay Gupta Alexander Schwing Derek Hoiem
University of Illinois at Urbana Champaign
{tgupta6, aschwing, dhoiem}@illinois.edu

The supplementary material includes:

1. Supervised Partitioning Analysis (Sec. 1)
2. Performance on all metrics for both unsupervised clustering and supervised partitioning analysis (Tab. 1)
3. Words with coarse and fine annotations used for clustering and partitioning analysis (Tab. 2, Sec. 3).
4. Categories used in the zero-shot analysis (Sec. 3)
5. Why are random vectors competitive with learned embeddings on vision-language tasks? (Sec. 4)

1. Supervised Partitioning Analysis

The partitioning analysis characterizes how well word embeddings represent the differences between words belonging to different semantic categories *while* sharing some representation with words in the same category. This is done by measuring the ability of a supervised learning algorithm to partition words into high-level categories at different learning capacities. Specifically, we use a decision tree classifier trained with Gini impurity as the splitting criterion and a minimum of 2 samples per leaf node. We control model capacity through maximum tree depth. We chose decision trees because: (i) they provide a natural way to control model capacity through depth, and (ii) they can hierarchically partition high dimensional spaces based on a label assignment.

For evaluation, in addition to accuracy, we also compute ARI and V-Measure on the induced clustering (words with the same predicted label belong to the same cluster). Note that we do not have a train-test split here since our goal is to study the separability of concepts in the embedding space rather than generalization.

Tab. 1 shows the average performance across tree depth for the 3 metrics. Our main conclusions from the partitioning analysis are as follows:

ViCo outperforms other embeddings. *ViCo* embeddings alone are partitioned better than other embeddings. *GloVe+ViCo* yields further improvements.

Coarse categories well represented in GloVe. While *GloVe+ViCo* yields significant gains (11 to 27% relative gain) over *GloVe* for fine classes, coarse categories are partitioned quite well with *GloVe* alone and using *ViCo* yields relatively small improvements (3 to 6% relative gain).

2. Consistency across tasks and metrics

Tab. 1 shows clustering and partitioning performance for multiple metrics – ARI and V-Measure for clustering, and Accuracy, ARI, and V-Measure for partitioning analysis. Our key conclusions are consistent across both tasks and metrics: (i) *ViCo* clusters words into visual categories better than other embeddings; (ii) Words in the *ViCo* embedding space are easier to partition into visual categories using a supervised learning algorithm; (iii) *GloVe+ViCo* outperforms all embeddings including *GloVe* and *ViCo* individually, showing the complementary nature of information encoded by the two embeddings.

3. Words and Categories

Tab. 2 shows the 495 words used in our clustering and partitioning analysis annotated with 13 coarse and 65 fine categories. The words were selected from the list of most frequent words in the VisualGenome [2] dataset, and were annotated manually with coarse and fine categories.

For the zero-shot analysis, we use CIFAR-100 [3]. The 100 categories are grouped into 20 super-categories consisting of 5 categories each. The super-categories are only used for generating the seen/unseen splits as described in the main submission (Sec.4). All categories and super-categories can be found at the original CIFAR-100 website: https://www.cs.toronto.edu/ kriz/cifar.html.

4. Why are random vectors competitive with learned embeddings?

Random vectors are surprisingly competitive with learned embeddings (both GloVe and ViCo) on vision-language tasks. Below, we present a hypothesis for this

Embeddings	Dim.	Unsupervised Clustering				Supervised Partitioning					
		Fine		Coarse		Fine			Coarse		
		V	ARI	V	ARI	V	ARI	Acc.	V	ARI	Acc.
random(100)	100	0.34	0.00	0.15	0.00	0.55	0.32	0.49	0.59	0.55	0.72
GloVe	300	0.50	0.15	0.52	0.38	0.70	0.48	0.64	0.77	0.74	0.84
GloVe+random(100)	300+100	0.50	0.14	0.49	0.35	0.70	0.48	0.65	0.76	0.74	0.84
vis-w2v-wiki [1]	200	0.41	0.08	0.43	0.27	0.73	0.52	0.67	0.77	0.74	0.84
vis-w2v-coco [1]	200	0.45	0.08	0.4	0.22	0.72	0.46	0.66	0.72	0.67	0.81
GloVe+vis-w2v-wiki	300+200	0.5	0.14	0.52	0.4	0.72	0.49	0.66	0.76	0.72	0.84
GloVe+vis-w2v-coco	300+200	0.52	0.16	0.55	**0.42**	0.74	0.56	0.68	0.77	0.74	**0.85**
ViCo(linear,100)	100	**0.60**	**0.21**	**0.59**	0.36	**0.76**	**0.57**	**0.70**	**0.78**	**0.76**	**0.85**
GloVe+ViCo(linear,100)	300+100	**0.61**	**0.23**	**0.65**	**0.48**	**0.78**	**0.61**	**0.72**	**0.81**	**0.78**	**0.87**

Table 1: **Comparing ViCo to other embeddings on clustering and partitioning tasks.** V-Measure for clustering is reported in the main submission. Here we show ARI for clustering, and V-Measure, ARI, and Accuracy for partitioning analysis. Conclusion are consistent across both tasks, and all metrics: (i) *ViCo* alone outperforms *GloVe*, *random*, and *vis-w2v* on all metrics, and their combinations on all but one metric (clustering ARI on coarse categories where *GloVe* and *GloVe+vis-w2v-** do better); (ii) *GloVe+ViCo* outperforms all other embeddings including *ViCo* and *GloVe*, showing that *ViCo* and *GloVe* encode complementary information.

behavior and test the hypothesis on image to caption retrieval task.

Hypothesis: *Given enough data, vision-language models learn to transform random vectors to get useful intermediate word representations.*

Test: Fig. 4 shows the performance of random and learned embeddings when trained on different amounts of training data. We see that learned embeddings have a significant advantage over random ones when the model is trained with only 1-2% of the available training data but diminishing gains (green line) are observed with more data.

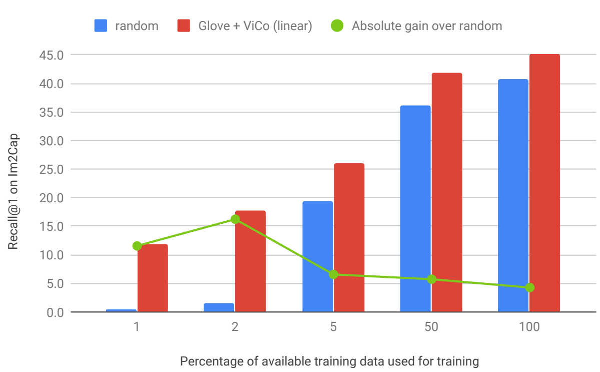

Figure 1. Comparing random and learned embeddings for Im2Cap model trained with varying amounts of data. We report average recall across 3 runs because of variance observed during training.

Reason for limited improvement of ViCo over Random and GloVe on VQA and Captioning. Because of the above hypothesis and availability of sufficient training data for tasks like VQA and Image Captioning, gains due to learned embeddings (for both GloVe and ViCo) are relatively small in comparison to random vectors.

However, we want to emphasize that our clustering, partitioning, and zero-shot analysis, as well as the discriminative attributes task highlight the advantages of learned embeddings over random embeddings, and ViCo over existing word embeddings. Finally, the ability to represent multiple senses of relatedness (Fig. 3 in the main submission) also distinguishes ViCo from existing word embeddings.

References

[1] Satwik Kottur, Ramakrishna Vedantam, José M. F. Moura, and Devi Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. *CVPR*, 2016. 2

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1

[3] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1

Coarse Categories	Fine Categories	Words
food	dessert	muffin, cake, pancake, sweet, candy, dessert, cupcake, doughnut, pastry, sugar
	drinks	coffee, tea, water, juice, beer, wine, alcohol, drink, milk
	fruits	apple, banana, fruit, pineapple, mango, pear, berry, lime, lemon, peach, plum, date
	herbs	oregano, herb, parsley, basil
	meals	lunch, meal, breakfast, dinner
	meats	pepperoni, steak, chicken, meat, pork
	nuts	nut, almond, cashew, pecan, peanut, walnut, hazelnut
	spices	salt, pepper, spice, chili, garlic, ginger
	vegetarian	tomato, zucchini, broccoli, vegetable, capsicum, spinach, onion, pea, squash, potato, corn, bean, cabbage, mushroom, carrot
	prepared/dishes	dish, chip, tortilla, burger, toast, bagel, pizza, pasta, sauce, salad, bread, pickle, bun, soup, noodle, syrup
	miscellaneous	dough, cheese, food, egg, wheat, rice, butter, oil
animals	birds	bird, turkey, owl, sparrow, pigeon, ostrich, duck, goose, swan, gull, flamingo, peacock
	farm	ox, cow, goat, cattle, bull, lamb, horse, donkey
	fish	fish, dolphin, shark
	pets	dog, cat, kitten, puppy
	reptiles	snake, reptile, turtle, crocodile, lizard
	wild	zebra, elephant, giraffe, lion, tiger, monkey, antelope, bear, animal, gazelle
colors	colors	red, green, blue, yellow, brown, grey, black, white, orange, purple, pink, cyan, violet, indigo, gold, silver, maroon
appliances	appliances	refrigerator, toaster, oven, burner, dishwasher, blender, microwave, oven, stove, appliance, cooler
electronics	computer	computer, laptop, keyboard, mouse, mousepad, printer
	display	monitor, television, tv, display
	audio	earbud, headphone, speaker, microphone
	communication	cellphone, phone, antenna, radio, telephone
	miscellaneous	electonic, device, digital, clock, camera, electronics
utensils	containers	cup, bowl, utensil, plate, jar, vase, urn
	drinks	glass, bottle, jug
	cutlery	spoon, spatula, fork
	cooking/brewing	pan, kettle, teapot, pot
transport	fourwheel	car, bus, minivan, tractor, buggy, van, minivan, jeep
	twowheel	scooter, bike, bicycle, motorcycle, moped
	rail	train, tram, railway, engine
	water	boat, ship, kayak
	air	aircraft, helicopter, jet, aeroplane, propeller
	generic	vehicle, transport, cargo
humans	profession	worker, firefighter, fireman, doctor, soldier, photographer, accountant, refree, student
	male	man, male, boy, father
	female	female, woman, girl, miss, mother, lady, girl
	neutral	people, person, individual, friend, lodger
	color	caucasian, brunet, blonde
	sport	skier, snowboarderskateboarder
	age	young, old, adolescent, teen, teenager, adult, child, baby
	commute	motorcyclist, cyclist, driver, bicyclist, motorist, pedestrian, rider
numbers	numbers	numeral, number, one, two, three, four, five, six, seven, eight, nine, ten
clothes	arms	wristband, mitten, glove, sleeve, watch
	coats	robe, blazer, jacket, coat
	full body	dress, apparel, suit, outfit, clothing, uniform, overall, full-dress
	head	sunglasses, spectacle, visor, helmet, beanie, cap, hat, headband, bandanna, hood
	legs	boot, sandal, shoe, slipper, legging, trouser, skirt, sock, jean, hosiery, stocking, pajama
	neck	bib, tie, scarf, necktie
	undergarments	swimsuit, bikini, lingerie, negligee
	torso	shirt, tshirt, t-shirt, top, blouse, apron, jersey, sweatshirt, sweater, lapel
construction	indoors	bedroom, room, stairway, hallway, patio, kitchen, bathroom, railing, balcony, ledge, lounge
	outdoors	ledge, balcony, roof, rooftop
	buildings	hotel, hostel, lodge, building, church, barn, shed, restaurant, church, hut
	commercial	booth, stall, market, plaza, shop, shopfront, mall
	fixtures	door, window, knob, faucet, lightbulb, bulb, latch, chandelier, fixture, sink, fireplace, bathtub
	furniture	chair, table, countertop, counter, cabinet, desk, bed, bench, cupboard, furniture, bookcase, armchair, sofa, bench, pew, stool, armchair, bookcase, seat, couch
actions	actions	call, take, step, come, get, bit, rid, throw, catch, enjoy, smile, eat, look, walk, stand, kneel, crouch, bend, talk, leave, construct, make, hit, keep, play, wait, relax, sit, read, serve, fix, lean, leave, kick, squat, bow, swing, get, go, gravel, annoy, rest, put, sleep, catch
bodyparts	face	ear, nose, nostril, head, lip, cheek, face, tounge, tooth, chin, thumb, elbow
	hair	hair, feather, tuft, eyebrow, brow, mane, eyelash, fur, ponytail, beard
	limbs	leg, arm, foot, thigh, calf, limb
	joints	knee, wrist, ankle, shoulder
	torso	thorax, waist, stomach, belly, abdomen, neck, hip
	extremeties	fingernail, finger, toe, toenail, hand
	non-human	tail, horn, claw, hoof

Table 2: **Words and Categories.** 495 words annotated with 13 coarse and 65 fine categories used in our clustering and partitioning analysis. Words were selected based on their frequency in VisualGenome and manually annotated with categories.