

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

1 Supplementary Material

1.1 Triplet Retrieval: More Examples

These are extended results of Figure 5 in the paper. We fix a dependency triplet and retrieve the image fragments that have the highest score with the triplet. These are sorted from left to right in decreasing score.

(AMOD, red, shorts)



(AMOD, blond, dog)



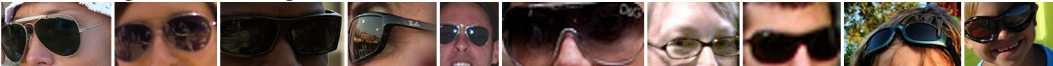
(DET, a, helmet)



(DET, a, rope)



(DOBJ, sunglasses, wearing)



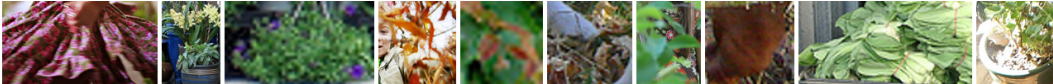
(NN, bike, rider)



(NN, swimming, trunks)



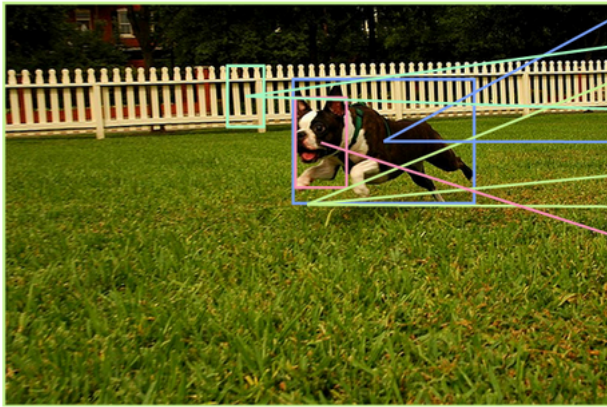
(PREP OF, flowers, bucket)



1.2 Image Annotation: More Examples

These are more examples of the results shown in Figure 4 in the paper. We fix an image and retrieve the top 5 most compatible sentences (shown below each image). The fragments of the top sentence are visualized next to the image and connected to the image fragment that they have the highest score with.

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107



- 67.6 (NSUBJ, terrier, running)
- 62.2 (DET, a, fence)
- 41.6 (PREP ON, grass, running)
- 33.6 (AMOD, white, fence)
- 29.4 (DET, a, terrier)
- 27.4 (AMOD, green, grass)
- 10.1 (AMOD, lush, grass)
- 0.2 (NN, boston, terrier)

1. A Boston Terrier is running on lush green grass in front of a white fence.
2. A German shepherd is showing its teeth as it growls at another German shepherd.
3. A dog runs on the green grass near a wooden fence.
4. A black and white dog is running in a grassy garden surrounded by a white fence.
5. A small dog jumps over a striped gate.



- 156.1 (DOBJ, sunglasses, wearing)
- 130.0 (NN, police, officer)
- 82.1 (NSUBJ, officer, smiles)
- 55.8 (PREP IN, uniform, officer)
- 52.1 (NN, cap, uniform)
- 30.1 (DET, a, officer)
- 17.8 (AMOD, female, officer)
- 13.9 (CONJ AND, navy, cap)
- 11.1 (DET, a, shop)
- 2.9 (DET, a, uniform)

1. A female police officer in a cap and navy uniform smiles while wearing sunglasses outside of a shop.
2. A female police officer, wearing an officer's hat and sunglasses, stands in uniform in front of a window-lined street block.
3. An african American police officer and woman stand at the front of a bus.
4. A child is wearing a blue knitted hat.
5. A police woman smiling and wearing sunglasses and a hat.

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161



- 82.8 (DET, a, sign)
- 49.6 (DET, a, cat)
- 49.5 (NSUBJ, cat, sits)
- 49.2 (NN, store, sign)

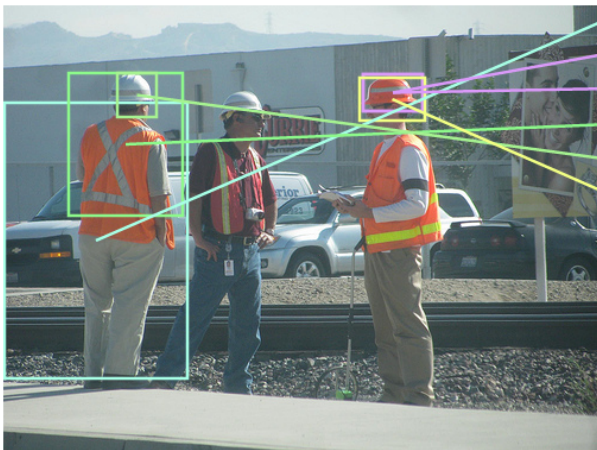
1. A cat sits on top of a store sign.
2. Man looking out the window of a train.
3. A cat sits upon a business sign.
4. A cat is sitting atop a sign on the side of a building.
5. Man standing at a counter in a convenience store with a woman on the other side of the counter wearing a white shirt looking at him very expectantly.



- 117.1 (AMOD, red, jacket)
- 50.3 (CONJ AND, hat, fur)
- 49.4 (NSUBJ, boy, looking)
- 47.0 (DET, a, hat)
- 42.4 (AMOD, blue, hat)
- 17.6 (PREP WITH, fur, wearing)
- 11.7 (AMOD, young, boy)
- 11.0 (VMOD, wearing, boy)
- 8.1 (PRT, down, looking)
- 6.4 (DET, a, boy)
- 4.1 (DOBJ, jacket, wearing)
- 3.3 (DET, a, jacket)

1. A young boy or girl wearing a red jacket with fur and a blue hat is looking down.
2. A young African american boy with red pants standing in front of a house.
3. Two boys wearing red jackets are digging shovels into the dirt, the smaller boy in the yellow boots has a mohawk.
4. A little boy dressed in red pants is standing in the street.
5. Little boy in red pants, no shirt, picking his teeth.

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215



- 119.2 (AMOD, orange, vests)
- 106.0 (NN, construction, workers)
- 43.8 (NSUBJ, workers, speak)
- 43.1 (PREP IN, vests, workers)
- 17.0 (NUM, three, workers)
- 1.1 (NN, railroad, tracks)

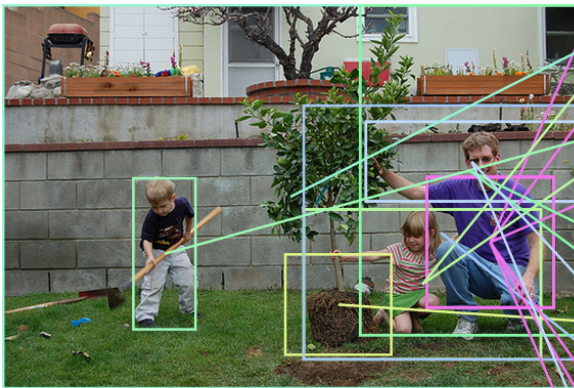
1. Three construction workers in orange vests speak to one another in front of railroad tracks.
2. Construction worker in orange vest and white hard hat with a shovel.
3. Several construction workers with orange safety vests are digging into the ground.
4. Construction workers in orange and yellow vests with jeans work on a street.
5. Construction workers picketing against PM Construction Services.

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269



- 107.6 (DOBJ, picture, taking)
- 57.7 (AMOD, digital, camera)
- 52.8 (DET, a, camera)
- 48.6 (PREP WITH, camera, taking)
- 30.3 (AMOD, tattooed, woman)
- 19.7 (VMOD, taking, woman)
- 11.1 (DET, a, picture)
- 6.9 (DET, a, woman)

1. A tattooed woman taking a picture with a digital camera.
2. Somebody took a photo of a girl with long black hair taking a photo.
3. A woman with a camera on a tripod is smiling for another camera.
4. A woman dressed in black with a tattoo on her right arm is taking a picture with her camera.
5. People enjoying a festival while having drinks and taking pictures on the cellphone.

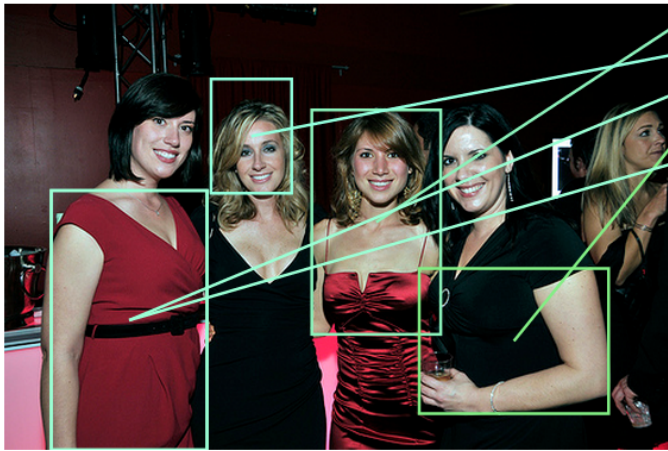


- 71.2 (NSUBJ, man, planting)
- 59.1 (DET, a, tree)
- 39.4 (AMOD, grassy, area)
- 38.9 (CONJ AND, girl, man)
- 32.7 (NSUBJ, boy, holding)
- 29.5 (PREP IN, area, girl)
- 29.1 (DET, a, hoe)
- 28.1 (AMOD, little, girl)
- 27.0 (AMOD, little, boy)
- 25.5 (DOBJ, hoe, holding)
- 21.4 (DOBJ, tree, planting)
- 8.4 (DET, a, girl)
- 7.9 (DET, a, boy)

1. A man and a little girl in a grassy area are planting a tree while a little boy off to the side is holding a hoe.
2. A father-figure and two children outside their home doing yard work such as using a hoe on the grass and planting a tree.
3. Three little children in a grassy yard running towards the camera.
4. A man and two children are planting a tree.
5. A woman is playing with two children on a seesaw in a playground.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

An issue with sentences as shown in this example is that the model does not reason about numbers and it is not natural to align triplets such as (NUM, two, individuals) to a detection. Also note the highly confident detection of camera on a person's face. These mistakes are due to people often describing frontal faces of people as "looking into the camera".



- 45.0 (NSUBJ, individuals, posing)
- 23.3 (DET, the, camera)
- 19.4 (PRT, up, dressed)
- 14.8 (VMOD, dressed, individuals)
- 8.8 (NUM, two, individuals)

1. Two individuals dressed up like animals are posing for the camera .

2. Four girls in evening attire pose for a picture .

3. two girls dress up for halloween .

4. Two girls wearing goth trendy clothing while one looks back at the camera .

5. four girls in evening wear are posing for a photograph .

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377

For the image below, we note that there are 18 mentions of “Miami” in the training set.



- 49.2 (NN, basketball, player)
- 45.8 (NN, miami, player)
- 9.4 (NSUBJ, player, looking)
- 2.8 (DET, the, player)

1. The Miami basketball player is looking .
2. A man in a Miami basketball uniform looking to the right
3. A girl with a nose ring , earring , and Mohawk .
4. A Miami basketball player dribbles by an Arizona State player .
5. A Miami basketball player looks off in the distance .

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

1.3 Image Search: Examples

In this experiment on Flickr8K test set we fix a sentence and retrieve the most compatible images. The images are ranked from left to right in decreasing confidence.

A black dog with brown on his face is swimming .



A black and white dog is looking at a pink Frisbee on a court .



A girl in a red jacket , surrounded by people .



The boy in the white jersey is dribbling the orange Wizards basketball .



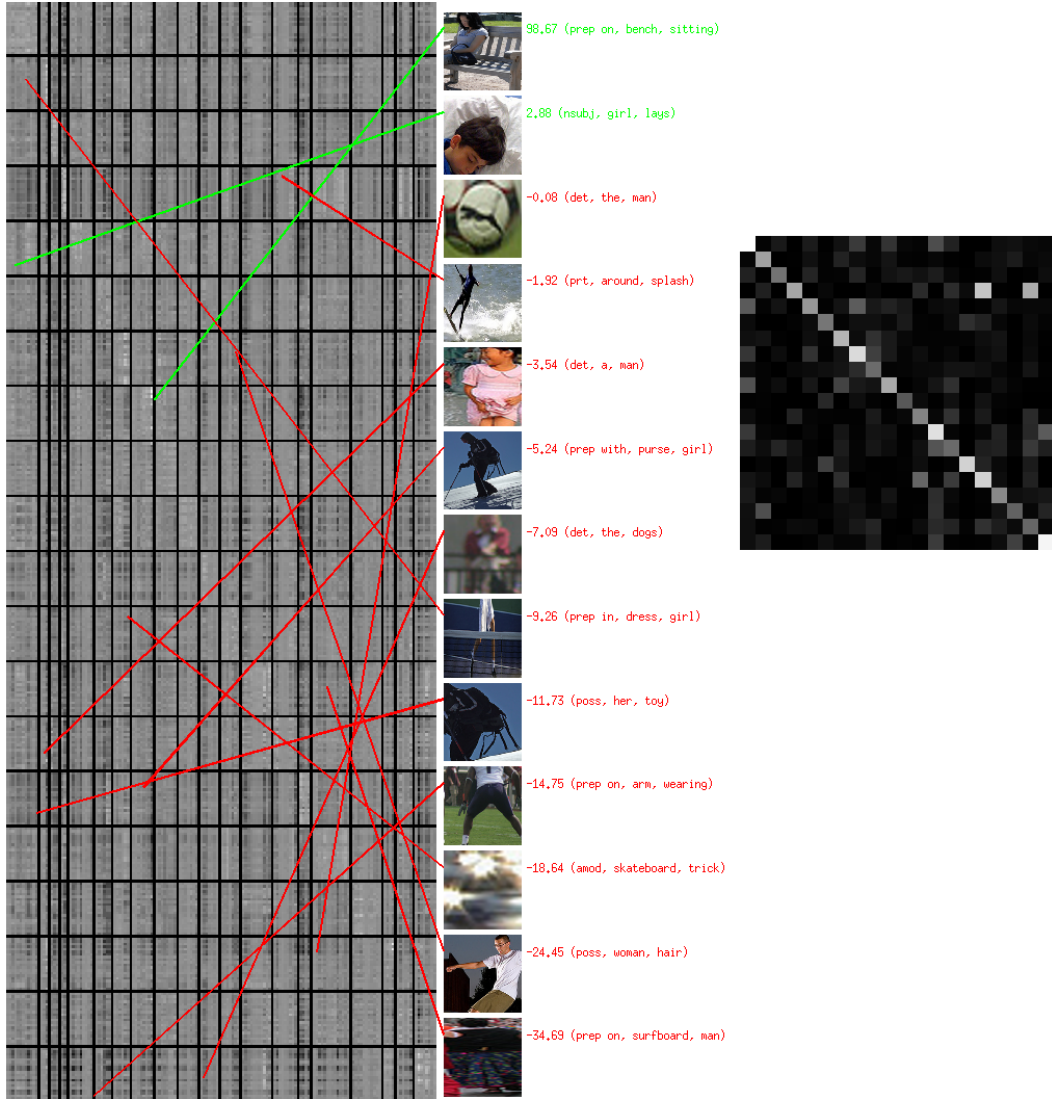
The person in the red and black uniform has a ball above her head .
(one of the current limitations is that the model does not reason about relative positions of objects)



432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

1.4 Score Matrix Visualization

We visualize some real examples of Figure 3 in the paper. Left: The raw fragment scores. Right: image-sentence scores for the batch of 20 images and sentences, once accumulated from fragment scores along blocks. The black grid lines delineate the boundaries between images and sentences. The red and green lines and images on side zoom in on particular examples of image-sentence fragment pairs from the matrix. Green indicates score greater than 0, red less than 0.



486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

