

Beyond a Pre-Trained Object Detector: Cross-Modal Textual and Visual Context for Image Captioning

Anonymous CVPR submission

Paper ID 4855

1. Proposed Graphical Model

In this section, we derive the graphical model with a newly introduced node T shown in Figure 1.

$$\begin{aligned}
 p(\mathbf{y}|\mathbf{x}) &= \prod_i p(y_i|\mathbf{x}, y_{1:i-1}) \\
 &= \prod_i \sum_{O,T} p(\mathbf{o}, \mathbf{t}|\mathbf{x}, y_{1:i-1}) p(y_i|\mathbf{x}, \mathbf{o}, \mathbf{t}, y_{1:i-1}) \\
 &= \prod_i \sum_{O,T} p(\mathbf{o}, \mathbf{t}|\mathbf{x}) p(y_i|\mathbf{x}, \mathbf{o}, \mathbf{t}, y_{1:i-1}) \quad (1) \\
 &= \prod_i \sum_{O,T} p(\mathbf{o}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) p(y_i|\mathbf{x}, \mathbf{o}, \mathbf{t}, y_{1:i-1}) \quad (2) \\
 &= \prod_i \sum_{O,T} p(\mathbf{o}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) p(y_i|\mathbf{o}, \mathbf{t}, y_{1:i-1}) \quad (3) \\
 &\simeq \prod_i \sum_T p(\mathbf{t}|\mathbf{x}) p(y_i|\mathbf{o}, \mathbf{t}, y_{1:i-1}) \quad (4) \\
 &\simeq \prod_i p(y_i|\mathbf{o}, \mathbf{t}, y_{1:i-1}) \quad (5)
 \end{aligned}$$

Between Equation 1 and Equation 2, given the input image \mathbf{x} , \mathbf{o} and \mathbf{t} are conditionally independent given the structure of the graphical model shown in Figure 1. Between Equation 2 and Equation 3, assume \mathbf{o} and \mathbf{t} completely encode all necessary information of \mathbf{x} , y_i is conditionally independent of \mathbf{x} . Between Equation 3 and Equation 4, researchers typically take argmax and threshold to select a fixed set of detected objects from the object detec-

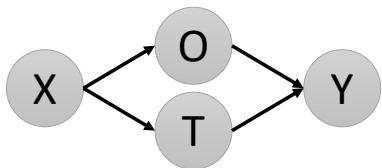


Figure 1. Graphical model with a newly introduced node T , which represents a set of retrieved text descriptions of image sub-regions.

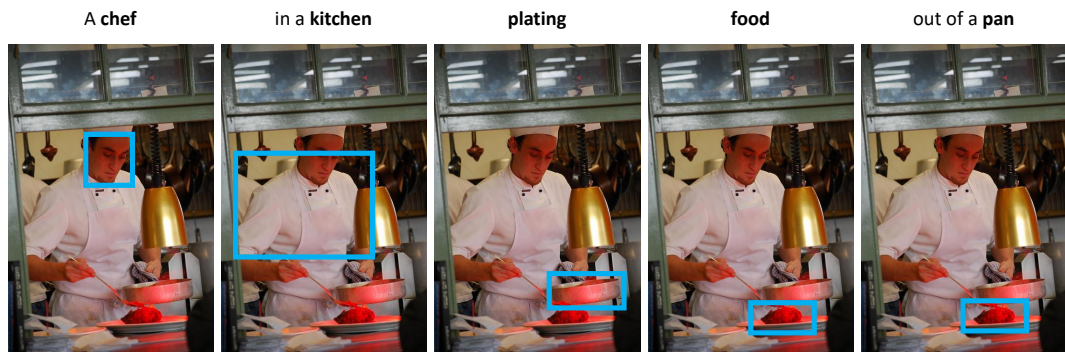
tor. Between Equation 4 and Equation 5, we propose to retrieve a fixed set of top- k most relevant text descriptions for each image crop. Eventually, we arrive at the same result as Equation 4 in the main paper.

2. Additional Qualitative Results

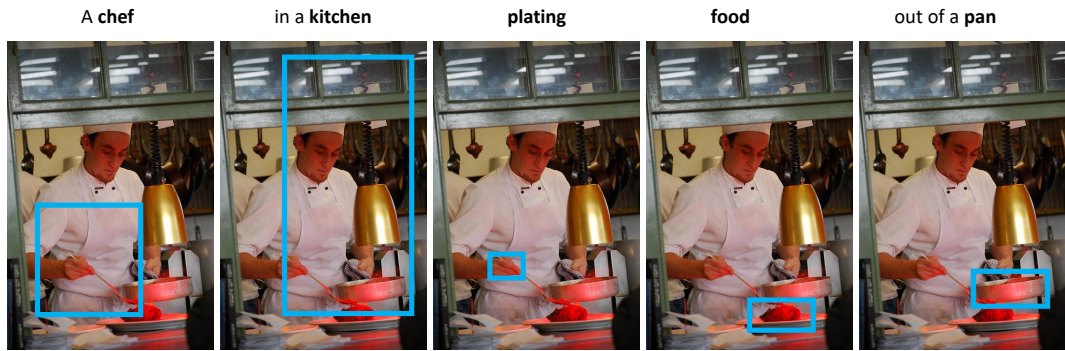
In Section 4.3 of the main paper, we quantitatively show that image conditioning helps refine the object features to aid with grounding on the Flickr30K dataset. In this section, we show qualitative examples in Figure 2-6 to further support this claim.

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

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

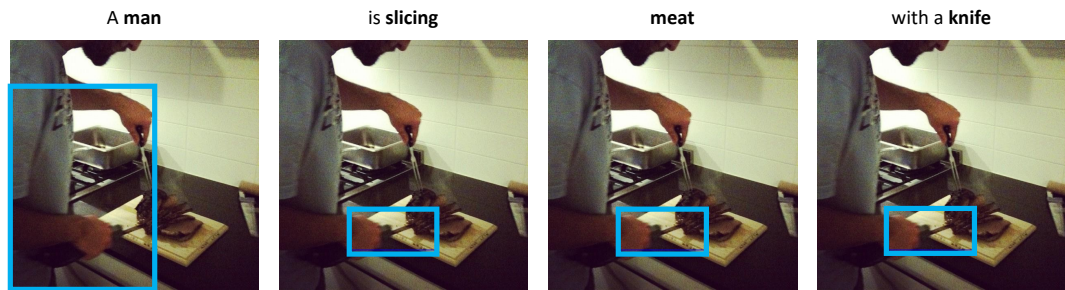


(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.

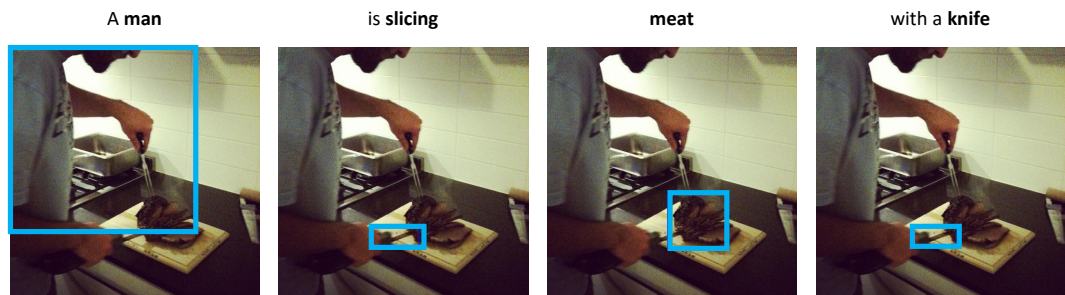


(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 2. Example of: *A chef in a kitchen plating food out of a pan.*



(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.

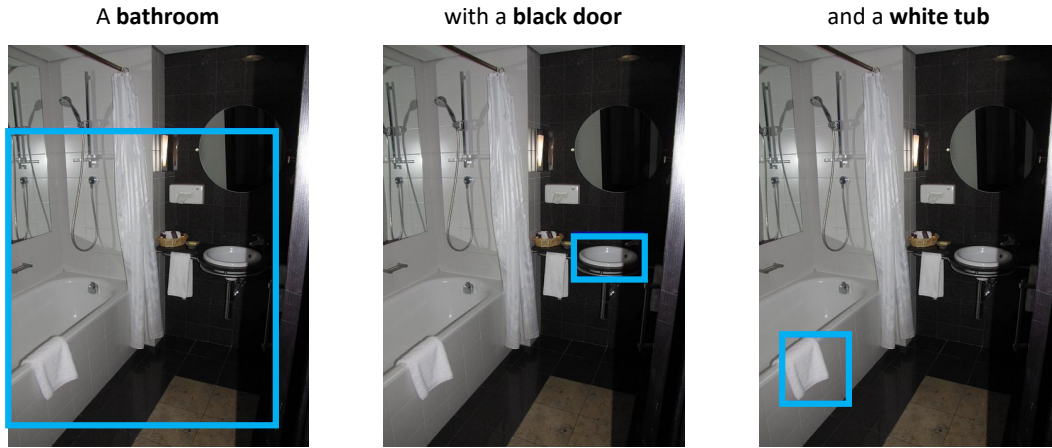


(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

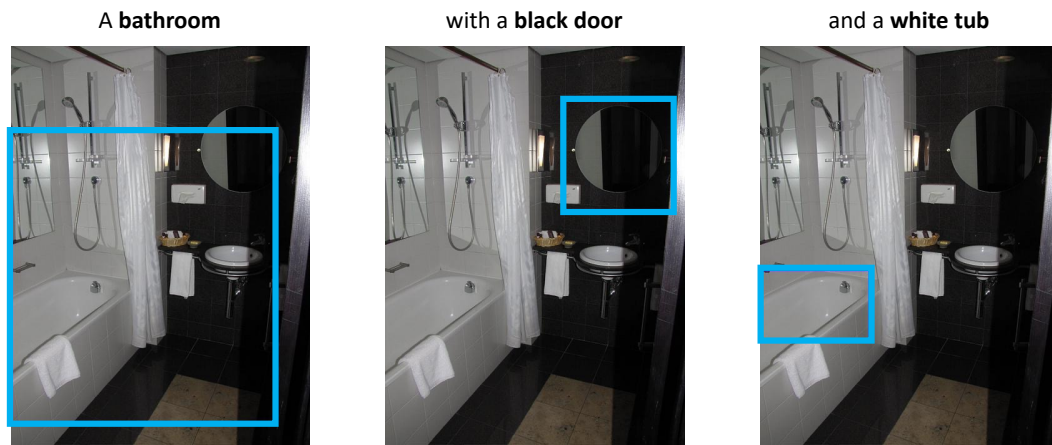
Figure 3. Example of: *A man is slicing meat with a knife.*

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

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



(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.



(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 4. Example of: *A bathroom with a black door and a white tub.*



(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.

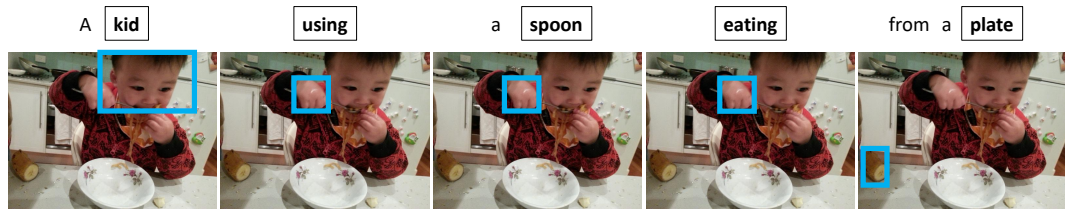


(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

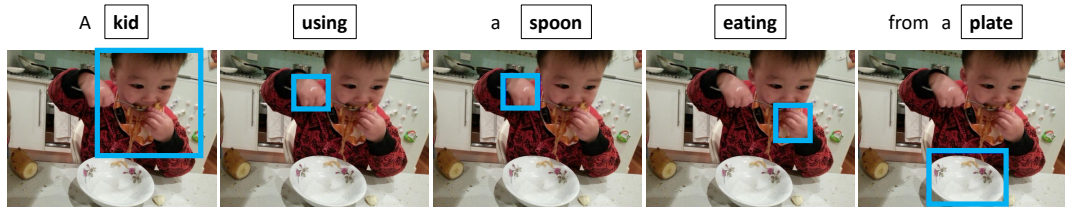
Figure 5. Example of: *A pair of men looking at a tablet on a table.*

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

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



(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.



(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 6. Example of: *A pair of men looking at a tablet on a table.*