Generating Image Descriptions with Gold Standard Visual Inputs: Motivation, Evaluation and Baselines (Errata)

Josiah Wang
Department of Computer Science
University of Sheffield
United Kingdom
j.k.wang@sheffield.ac.uk

Robert Gaizauskas
Department of Computer Science
University of Sheffield
United Kingdom
r.gaizauskas@sheffield.ac.uk

Errata

There was a bug in our original implementation of the visual prior based on the **positions of bounding boxes**. The correct Precision/Recall/F scores for this particular baseline are in actual fact much lower than reported in the paper, and lower than all proposed baselines (except the random baseline). As such, we infer that bounding box position may be a weaker visual cue compared to bounding box size, at least for this particular dataset.

This document shows the corrected results.

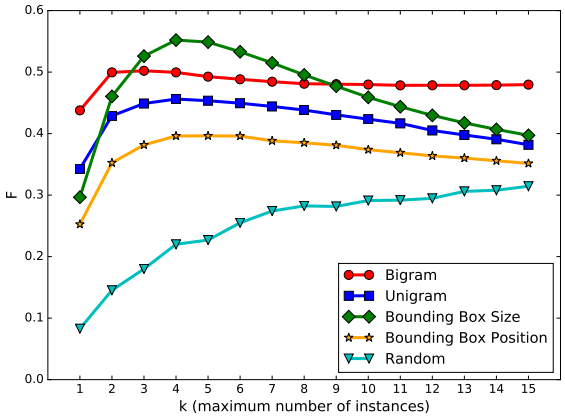


Figure 1: Replaces **Figure 3** of the original paper. The figure shows the content selection score, F, evaluated on the proposed baselines at varying levels of k (maximum number of instances per sentence). Standard deviations are omitted for clarity, but are included in Table 1.

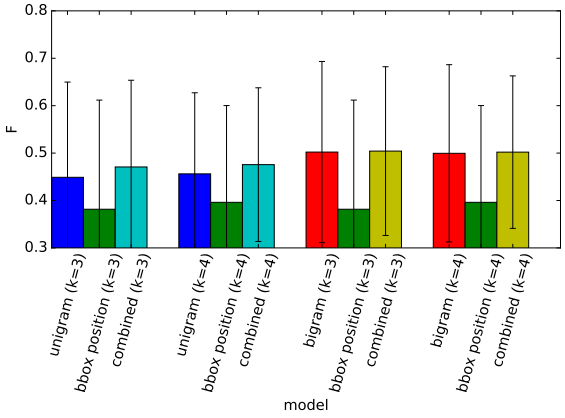


Figure 2: Replaces **Figure 6** of the original paper. The figure shows the content selection score, F, when combining textual priors (unigram or birgram) and visual cues based on **bounding box positions**. We compare the combined baselines at k=3 and k=4.

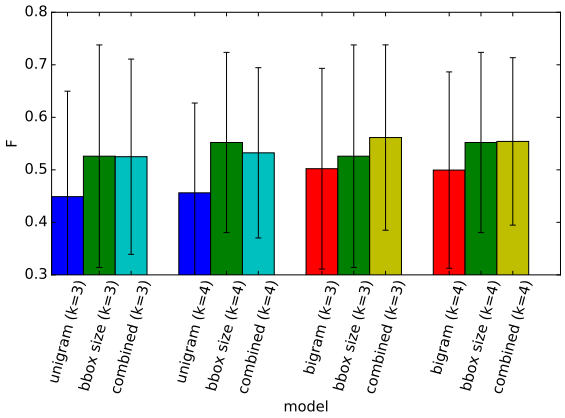
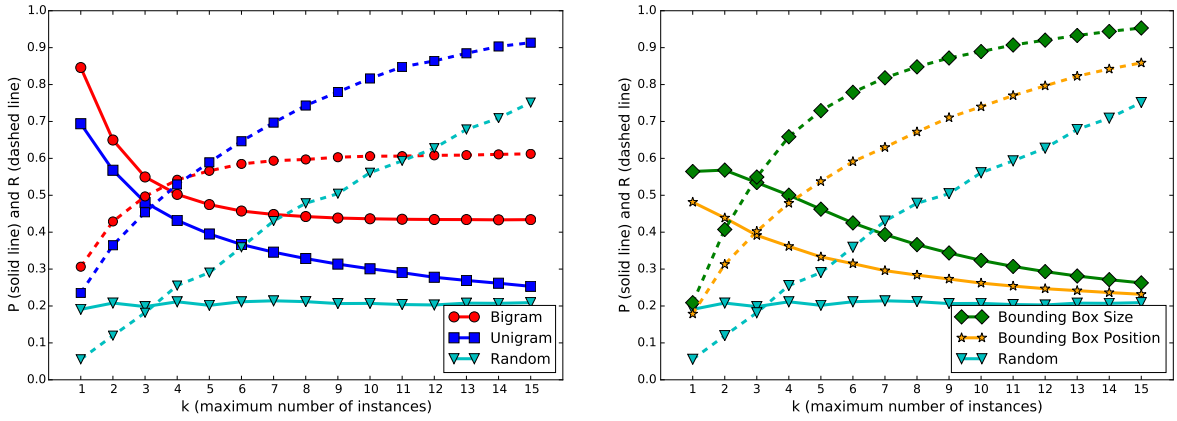


Figure 3: The content selection score, F, when combining textual priors (unigram or birgram) and visual cues based on **bounding box sizes**. We compare the combined baselines at k=3 and k=4. This figure is provided as a supplement, as our initial claim that using bounding box position and using bounding box size yield similar results does not now hold.

		P	R	F
Random	$k=1$	0.19 ± 0.32	0.06 ± 0.12	0.08 ± 0.16
	$k=2$	0.21 ± 0.24	0.12 ± 0.17	0.15 ± 0.19
	$k=3$	0.20 ± 0.20	0.18 ± 0.22	0.18 ± 0.20
	$k=4$	0.21 ± 0.18	0.26 ± 0.25	0.22 ± 0.20
	$k=5$	0.20 ± 0.17	0.29 ± 0.27	0.23 ± 0.19
	$k=6$	0.21 ± 0.17	0.36 ± 0.29	0.25 ± 0.19
	$k=7$	0.21 ± 0.15	0.43 ± 0.31	0.27 ± 0.18
	$k=8$	0.21 ± 0.15	0.48 ± 0.31	0.28 ± 0.18
	$k=9$	0.21 ± 0.15	0.50 ± 0.32	0.28 ± 0.18
	$k=10$	0.21 ± 0.14	0.56 ± 0.31	0.29 ± 0.17
Bounding Box Position	$k=1$	0.48 ± 0.43	0.18 ± 0.20	0.25 ± 0.26
	$k=2$	0.44 ± 0.29	0.31 ± 0.25	0.35 ± 0.26
	$k=3$	0.39 ± 0.22	0.40 ± 0.27	0.38 ± 0.23
	$k=4$	0.36 ± 0.18	0.48 ± 0.28	0.40 ± 0.20
	$k=5$	0.33 ± 0.16	0.54 ± 0.28	0.40 ± 0.18
	$k=6$	0.31 ± 0.14	0.59 ± 0.27	0.40 ± 0.16
	$k=7$	0.30 ± 0.14	0.63 ± 0.27	0.39 ± 0.16
	$k=8$	0.28 ± 0.14	0.67 ± 0.27	0.39 ± 0.15
	$k=9$	0.27 ± 0.13	0.71 ± 0.26	0.38 ± 0.15
	$k=10$	0.26 ± 0.13	0.74 ± 0.25	0.37 ± 0.14
Bounding Box Size	$k=1$	0.56 ± 0.41	0.21 ± 0.19	0.30 ± 0.25
	$k=2$	0.57 ± 0.27	0.41 ± 0.25	0.46 ± 0.25
	$k=3$	0.53 ± 0.20	0.55 ± 0.26	0.53 ± 0.21
	$k=4$	0.50 ± 0.16	0.66 ± 0.24	0.55 ± 0.17
	$k=5$	0.46 ± 0.14	0.73 ± 0.22	0.55 ± 0.15
	$k=6$	0.43 ± 0.14	0.78 ± 0.20	0.53 ± 0.13
	$k=7$	0.39 ± 0.13	0.82 ± 0.19	0.51 ± 0.12
	$k=8$	0.37 ± 0.13	0.85 ± 0.17	0.50 ± 0.12
	$k=9$	0.34 ± 0.13	0.87 ± 0.16	0.48 ± 0.12
	$k=10$	0.32 ± 0.12	0.89 ± 0.15	0.46 ± 0.12
Unigram	$k=1$	0.69 ± 0.40	0.24 ± 0.18	0.34 ± 0.24
	$k=2$	0.57 ± 0.29	0.36 ± 0.22	0.43 ± 0.23
	$k=3$	0.48 ± 0.22	0.45 ± 0.23	0.45 ± 0.20
	$k=4$	0.43 ± 0.19	0.53 ± 0.23	0.46 ± 0.17
	$k=5$	0.40 ± 0.17	0.59 ± 0.22	0.45 ± 0.16
	$k=6$	0.37 ± 0.16	0.65 ± 0.22	0.45 ± 0.15
	$k=7$	0.35 ± 0.15	0.70 ± 0.22	0.44 ± 0.14
	$k=8$	0.33 ± 0.14	0.74 ± 0.22	0.44 ± 0.14
	$k=9$	0.31 ± 0.14	0.78 ± 0.21	0.43 ± 0.14
	$k=10$	0.30 ± 0.13	0.82 ± 0.20	0.42 ± 0.13
Bigram	$k=1$	0.85 ± 0.29	0.31 ± 0.17	0.44 ± 0.21
	$k=2$	0.65 ± 0.24	0.43 ± 0.21	0.50 ± 0.21
	$k=3$	0.55 ± 0.21	0.50 ± 0.22	0.50 ± 0.19
	$k=4$	0.50 ± 0.21	0.54 ± 0.23	0.50 ± 0.19
	$k=5$	0.47 ± 0.21	0.57 ± 0.23	0.49 ± 0.19
	$k=6$	0.46 ± 0.20	0.59 ± 0.23	0.49 ± 0.18
	$k=7$	0.45 ± 0.20	0.59 ± 0.23	0.48 ± 0.18
	$k=8$	0.44 ± 0.20	0.60 ± 0.23	0.48 ± 0.18
	$k=9$	0.44 ± 0.20	0.60 ± 0.24	0.48 ± 0.18
	$k=10$	0.44 ± 0.20	0.61 ± 0.23	0.48 ± 0.18

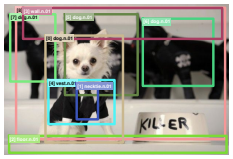
Table 1: Replaces **Table 1** of the original paper. The table shows the P, R and F scores (with standard deviations) of the content selection metric, as evaluated on different baselines at varying levels of k (1 to 10).



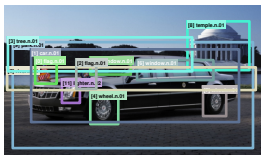
(a) Baselines based on textual priors

(b) Baselines based on visual cues

Figure 4: Replaces **Figure 4** of the original paper. The precision P (solid lines) and recall R (dashed lines), as evaluated on the proposed baselines at varying levels of k. Again, error bars are omitted for clarity, but are included in Table 1.

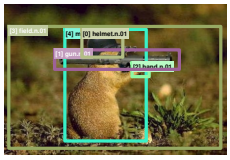


random:	[F=0.04]	[Wall]^{3} among [necktie]^{1} underneath [floor]^{2} .
bbox pos:	[F=0.45]	[Hallway]^{8} below the [dog]^{5} near the [dog]^{0} .
bbox size:	[F=0.39]	[Hallway]^{8} behind the [dog]^{0} underneath the [wall]^{3} .
unigram:	[F=0.05]	[Wall]^{3} near [floor]^{2} with the [dog]^{5} .
bigram:	[F=0.51]	[Dog]^{5} against [dog]^{0} beside the [dog]^{6} .

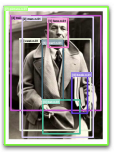


random:	[F=0.05]	[Park]^{9} behind [wheel]^{7} underneath the [window]^{6} .
bbox pos:	[F=0.59]	[River]^{5} on the [car]^{1} below [park]^{9} .
bbox size:	[F=0.44]	[Park]^{9} behind the [car]^{1} against the [tree]^{3} .
unigram:	[F=0.42]	[Tree]^{3} beneath [car]^{1} by [window]^{6} .
bigram:	[F=0.71]	[Car]^{1} inside [flag]^{0} underneath the [flag]^{2} .

random:	[F=0.43]	[Wall]^{4} inside [door]^{3} around the [bicycle]^{0} .
bbox pos:	[F=0.79]	[Bicycle]^{0} in [wall]^{2} below [floor]^{1} .
bbox size:	[F=0.79]	[Bicycle]^{0} on [floor]^{1} with [wall]^{2} .
unigram:	[F=0.34]	[Table]^{7} in the [wall]^{4} around [wall]^{2} .
bigram:	[F=0.03]	[Table]^{7} near [door]^{3} .



random:	[F=0.66]	[Mouse]^{4} inside [field]^{3} against [helmet]^{0} .
bbox pos:	[F=0.75]	[Field]^{3} and [mouse]^{4} beside the [gun]^{1} .
bbox size:	[F=0.75]	[Field]^{3} along [mouse]^{4} underneath [gun]^{1} .
unigram:	[F=0.31]	[Field]^{3} inside [hand]^{2} below [helmet]^{0} .
bigram:	[F=0.00]	[Hand]^{2} .



random:	[F=0.39]	[Vest]^{6} at [hat]^{3} behind the [picture]^{7} .
bbox pos:	[F=0.41]	[Picture]^{7} on [man]^{2} beside the [scarf]^{5} .
bbox size:	[F=0.49]	[Picture]^{7} among [man]^{2} on the [train]^{4} .
unigram:	[F=0.77]	[Man]^{2} below the [hat]^{3} at [book]^{0} .
bigram:	[F=0.77]	[Man]^{2} around the [hat]^{3} along the [book]^{0} .

Figure 5: Replaces **Figure 5** of the original paper. Example image descriptions generated by our baselines ($k = 3$).