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

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

Abstract

In this paper, we present the task of generating image descriptions with gold standard visual detections as input, rather than directly from an image. This allows the Natural Language Generation community to focus on the text generation process, rather than dealing with the noise and complications arising from the visual detection process. We propose a fine-grained evaluation metric specifically for evaluating the *content selection* capabilities of image description generation systems. To demonstrate the evaluation metric on the task, several baselines are presented using bounding box information and textual information as priors for content selection. The baselines are evaluated using the proposed metric, showing that the fine-grained metric is useful for evaluating the content selection phase of an image description generation system.

1 Introduction

There has been increased interest in the task of automatically generating full-sentence natural language image descriptions in recent years. Compared to earlier work that annotates images with isolated concept labels (Duygulu et al., 2002), such detailed annotations are much more informative and discriminating, and are important for improved text and image retrieval. They also pose an interesting and difficult challenge for natural language generation.

Previous work on generating image descriptions concentrates on solving the problem 'end-to-end', that is to generate a description given an image as input (Yao et al., 2010; Kulkarni et al., 2011; Yang et al., 2011). Recent advances in large scale visual object recognition, especially in deep learning techniques, have reached a reasonably high level of accuracy in the last few years. For the task of classifying an image into one of 1,000 object categories (i.e. does the image contain an object of category X, yes or no?) on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC'14) dataset (Russakovsky et al., 2014), the state-of-the-art currently performs at a 4.82% top-5 error rate (Ioffe and Szegedy, 2015) comparable to the 5.1% error rate of a human annotator who trained himself to recognise the object categories (Russakovsky et al., 2014). For the more challenging object category detection task (i.e. draw a bounding box around each instance of objects of the given categories), the state-of-the-art achieved a mean average precision of 43.9%. However, even at this level of performance, the errors from the visual output are still problematic when used as input to an image description generation system, especially when considering a large pool of candidate object categories to be mentioned in the description.

What if we were to assume that visual object recognisers have already achieved close to perfect detection rates, and that the object instances have already been identified and localised in an image? This then raises many interesting questions with regards to generating a description for an image, including: (i) how do we decide which objects are to be mentioned? (ii) how should these objects be ordered in the description? (iii) how do we infer and describe activities or actions? (iv) how to we describe spatial relations between objects? (v) how and when do we describe the object attributes? Many of these questions could be explored if we had a 'perfect' visual input to our image description generator.

To be able to begin to answer these questions, we proposed a pilot task, which has formed part of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task bench-

This is the first author's self-archived version of the paper, posted on 6th April 2016 at http://www.josiahwang.com/. It includes an errata that corrects the results of the baseline based on bounding box positions.

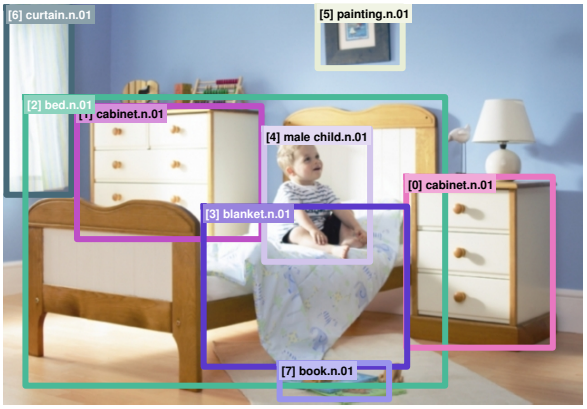


Figure 1: We present the task of generating textual descriptions given gold standard labelled bounding boxes as input. This allows researchers to focus on the text generation aspects of the image description generation task, rather than dealing with the noise arising from visual detection. This task also allows us to evaluate specific phases of the conventional generation pipeline, providing insights into which specific phases of the generation pipeline contribute to the performance of an image description generation system.

marking challenge (Villegas et al., 2015; Gilbert et al., 2015). More specifically, we assume that perfectly labelled object instances and their localisations are available to image description generation systems, as done in Elliott and Keller (2013) and Yatskar et al. (2014). Given this knowledge, we would like to evaluate how well image description generation systems perform through the various stages of Natural Language Generation (Reiter and Dale, 2000): content determination (what objects to describe), microplanning (how to describe objects) and realisation (generating the complete sentence). This pilot task is an attempt at encouraging fine-grained evaluation specifically for image descriptions, compared to general-purpose metrics like METEOR (Denkowski and Lavie, 2014) that evaluates text at a global, coarse-grained level. For our pilot, we concentrated on just one fine-grained metric: a content selection measure to evaluate how well a text generation system selects the correct object instances to be mentioned in the resulting image description.

A dataset has been introduced for this particular challenge. This paper will not discuss in great detail how the dataset has been collected and annotated; we instead refer readers to Gilbert et al. (2015) for more details about the challenge. The main purpose of this paper, instead, is to: (i) present and discuss the task of generating image descriptions with a *gold standard visual input*; (ii) propose a fine-grained metric specifically for evaluating the *content selection* capabilities of image description generation systems; (iii) introduce several *baselines* for this task and evaluate the baselines using the proposed fine-grained metric.

Overview. In section 2, we discuss the motivations for introducing the pilot task and the fine-grained metric in the ImageCLEF 2015 challenge, positioning them in relation to existing work. In section 3, we describe the task of generating image descriptions given gold standard visual inputs, along with a discussion on evaluating image description generation systems with regards to their content selection abilities. Section 4 presents several baselines for this task, while section 5 evaluates these baselines using the proposed content selection metric. Finally, we discuss further challenges with the proposed task, and introduce possible fine-grained metrics to be considered in the future.

2 Motivation and Related Work

There are currently three main groups of approaches to generating image descriptions. The most common and intuitive paradigm is the knowledge-based, generative approach that takes an image as input, detects instances of pre-defined object categories in the image using a visual recogniser, and then reasons about the detected objects to generate a novel textual description (Yao et al., 2010; Kulkarni et al., 2011; Yang et al., 2011; Li et al., 2011; Mitchell et al., 2012). However, these approaches are constrained to a limited number of categories, for example 20 in Kulkarni et al. (2011). We found that these approaches are generally sensitive to errors from visual input detection, as such errors tend to propagate and accumulate through the generation pipeline. The problem is accentuated when scaling up to a larger number of categories (e.g. 1000), where it becomes difficult to reason about what to describe amongst the candidate instances produced by the noisy visual detectors. Thus, generating image descriptions with gold standard visual input allows researchers to concentrate on the sentence generation aspects without being bogged down by the complications of the vision aspects of the task.

The second group of work revolves around de-

A [woman]^{2} in a white [dress]^{0} and gold [boots]^{5} leaning on a [car]^{3} .

A [woman]^{2} poses along a [car]^{3} .

[woman]^{2} dressed in white with gold [boots]^{5} poses next to a police [car]^{3}.

A [woman]^{2} dressed in white leans against a white [car]^{3} .

A [woman]^{2} is leaning against a [car]^{3}.

A [woman]^{2} with gold [boots]^{5} leans against an Indy pace [car]^{3} .

A blonde [woman]^{2} wearing gold shiny [boots]^{5} , a white [top]^{0} and short white skirt is leaning on a [car]^{3} .

Figure 2: An example image and its seven corresponding textual descriptions from the development dataset, with bounding box annotations labelled with WordNet concepts, and the correspondence of bounding boxes to entity mentions in the descriptions. For example, [woman]^{2} in the first sentence refers to bounding box ID [2] in the image, and [dress]^{0} corresponds to bounding box ID [0]. Correspondence is annotated at word level rather than at phrase level to avoid possible complications with multiple correspondences within the same phrase (*woman in a white dress*).

scription generation by retrieving existing textual descriptions from similar images. A common approach would be to map text and images to a common meaning space (Farhadi et al., 2010; Hodosh et al., 2013; Socher et al., 2014) or by using some similarity measure (Ordonez et al., 2011). Although such methods produce descriptions that are more expressive, they rely on a large amount of training data, and are unable to produce novel sentences. There have been attempts at retrieving only text fragments and combining them to generate novel descriptions (Kuznetsova et al., 2012; Kuznetsova et al., 2014) or by pruning irrelevant fragments for better generalisation (Kuznetsova et al., 2013). However, the resulting descriptions may still be pure 'guesswork' and may reference text fragments that are irrelevant to image content.

Most recently, work using deep learning approaches has produced state-of-the-art results (Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Vinyals et al., 2015), by utilising Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012; Razavian et al., 2014) as image features, and a Recurrent Neural Network (RNN) (Mikolov et al., 2010) for language modelling, and learning to generate descriptions jointly from images and their descriptions. The advantages of such models are that they cope better with noisy visual detections, and that the RNN language models are capable of modelling long range dependencies. The main disadvantages are (i) it is difficult to inspect what has been learnt by the model and hence to gain insight into what is working or not working in the system; (ii) these methods are dependent on image datasets aligned with sentences as learning is performed in a joint manner. The latter limitation means new datasets need to be produced even for small changes in the task, such as generating descriptions that are more or less detailed, or in more or less simplified language (e.g. for children) or have a specific information focus (say, focussing on buildings versus people in an image for a particular application). Thus, knowledge-based, generative approaches may have an advantage in this respect, as there is no need for aligned image-text datasets, since visual detection and sentence generation are independent, allowing the language model to be tuned at surface realisation stages.

Image description generation with gold standard input. As discussed, knowledge-based, generative approaches are sensitive to visual detection input errors. Therefore, previous work has proposed circumventing the problem by providing gold standard annotations as input to description generation systems. Elliott and Keller (2013) provide region annotations along with spatial relations between region instances. Yatskar et al. (2014) also provide gold standard region anno-

tations, as well as fine-grained region properties such as attributes, parts, and activities. Zitnick and Parikh (2013) take a unique approach of generating scenes from clipart as an abstraction to real world images to address the issue of noisy input. Our work takes a similar direction as Elliott and Keller (2013) and Yatskar et al. (2014), but with bounding boxes as gold standard input, and with an emphasis on fine-grained evaluation of image description generation systems.

Evaluation of image description generation systems. Existing image description generation systems are most commonly evaluated using automatic evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Denkowski and Lavie, 2014) and most recently CIDEr (Vedantam et al., 2015). However, such global measures only allow evaluation of image description generation systems as a whole, without being able to ascertain which parts of the generation process, or components of the generation system, are responsible for performance gains or losses. Although evaluations based on human judgments could provide a more fine-grained metric (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012), they are expensive and difficult to scale. We propose instead to exploit the pipeline of knowledge-based, generative approaches to generation, allowing us to inspect specific capabilities of image generation systems by means of evaluation with fine-grained metrics. Rather than just evaluating image description generation extrinsically with a global evaluation measure, we isolate evaluation of different phases of description generation, and treat each phase as a first-class citizen.

3 Task and Evaluation Measure

As mentioned above, we introduced as a benchmarking challenge the task of generating image descriptions for 450 test images given gold standard, labelled bounding box annotations as input (Figure 1). The category labels were restricted to 251 WordNet synsets selected specifically for the challenge. To enable evaluation with our proposed fine-grained metric, participants were also asked to annotate, within their generated descriptions, the bounding box ID to which a term in the description corresponds. A development dataset of 500 images was provided with labelled bounding box annotations and correspondence annotations between textual terms and bounding boxes. Figure 2 shows an example annotation of bounding boxes and the correspondences between bounding box instances and terms in the image descriptions. Note that correspondence was annotated at word level (unigram) rather than at phrase level (higher-order n-grams) to avoid possible complications with multiple correspondences within the same phrase (*woman in a white dress*).

3.1 Fine-grained Evaluation Metric

As a pilot, we propose a fine-grained metric to evaluate the content selection capabilities of an image description system. This *content selection* metric is the F_1 score averaged across all 450 test images, where each F_1 score is computed from the precision and recall averaged over all gold standard descriptions for the image.

Formally, let $I = \{I_1, I_2, ...I_N\}$ be the set of test images. Let $G^{I_i} = \{G_1^{I_i}, G_2^{I_i}, ..., G_M^{I_i}\}$ be the set of gold standard descriptions for image I_i, where each $G_m^{I_i}$ is the set of unique bounding box instances referenced in gold standard description m of image I_i. Let S^{I_i} be the set of unique bounding box instances referenced by the participant's generated sentence for image I_i. The precision P^{I_i} for test image I_i is computed as:

$$P^{I_i} = \frac{1}{M} \sum_{m}^{M} \frac{|G_m^{I_i} \cap S^{I_i}|}{|S^{I_i}|} \quad (1)$$

where $|G_m^{I_i} \cap S^{I_i}|$ is the number of unique bounding box instances referenced in both the gold standard description and the generated sentence, and M is the number of gold standard descriptions for image I_i.

Similarly, the recall R^{I_i} for test image I_i is computed as:

$$R^{I_i} = \frac{1}{M} \sum_{m}^{M} \frac{|G_m^{I_i} \cap S^{I_i}|}{|G_m^{I_i}|} \quad (2)$$

The content selection score for image I_i, F^{I_i}, is computed as the harmonic mean of P^{I_i} and R^{I_i}:

$$F^{I_i} = 2 \times \frac{P^{I_i} \times R^{I_i}}{P^{I_i} + R^{I_i}} \quad (3)$$

The final P, R and F scores are computed as the mean P, R and F scores across all test images.

The advantage of the macro-averaging process in equations (1) and (2) is that it implicitly captures the relative importance of the bounding box

instances based on how frequently they occur across the gold standard descriptions. For example, in Figure 2, both *woman* and *car* are referenced in all seven gold standard descriptions, while *boot* is mentioned four times and *dress* twice. Thus, a generated description that references *woman* and *car* will naturally result in a higher score than one that references only *woman* and *dress*.

Note that for this metric, we are only concerned with evaluating the generation system's *content selection* capabilities, rather than its referring expression generation. As such, systems are free to generate any referring expression for each selected bounding box instance. We consider the evaluation of referring expressions as a potentially separate fine-grained evaluation task to be introduced in the future. In addition, we do not evaluate terms outside those that refer to bounding box instances, and as for the pilot task of the challenge use the global METEOR metric to cover evaluation of other aspects of image description generation.

4 Generating Descriptions: Baselines

We propose a set of baselines for the image description generation task, or more specifically the content selection task. These allow us to test the proposed fine-grained content selection metric (Section 3.1) and to gain some insights into what features might inform content selection. The baselines use visual and textual cues to select the bounding box instances to be described in the text to be generated.

4.1 Generation based on Visual Cues

Stratos et al. (2012) found that the size and position of visual entities in an image, to a certain extent, plays a part in determining what is mentioned in the corresponding description. As such, we consider two baselines based on different visual cues: (i) bounding box size (bigger objects have higher likelihood of being mentioned); (ii) distance of the centroid of the bounding box to the centre of the image (central objects have higher likelihood of being mentioned). For each test image, bounding boxes instances are sorted based on these visual cues, and a fixed threshold used to limit the number of instances to be selected for sentence generation. We will explore different thresholds in our experiments in Section 5.

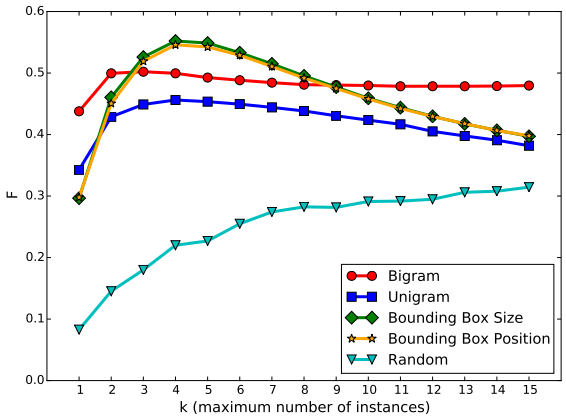


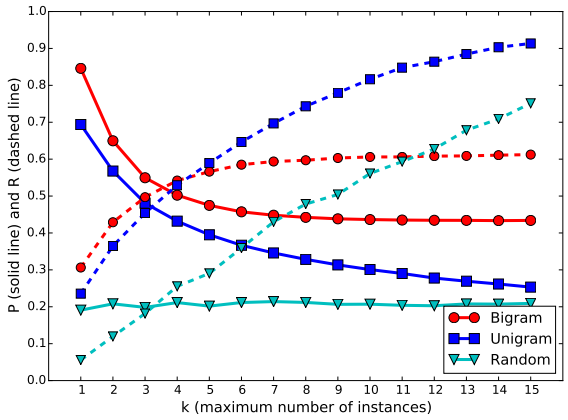
Figure 3: The content selection score, F, evaluated on the proposed baselines at varying levels of k (maximum number of instances per sentence). Standard deviations are omitted for clarity, but are included in Table 1.

4.2 Generation based on Textual Priors

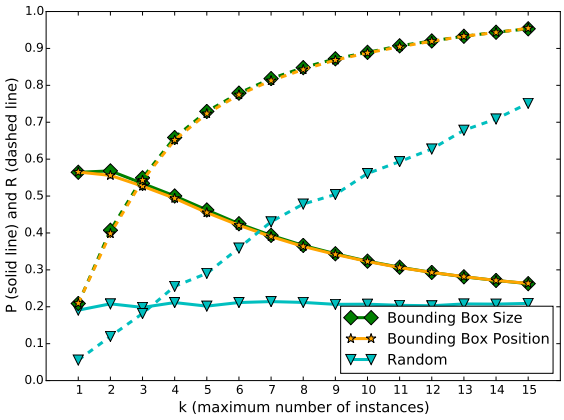
We also consider baselines based on textual priors, as Stratos et al. (2012) also showed that the category of the object play a role in determining whether it will be mentioned in the corresponding textual description.

For the first baseline, we consider as a prior unigram counts of concepts that have been referenced to a bounding box in the gold standard descriptions from the development set. For each test image, bounding boxes are sorted by the frequency of their concept labels in the development set, i.e. frequently mentioned concepts have higher precedence.

We also consider a more sophisticated baseline based on bigram sequences, where a concept is selected based on how likely it is to be referenced *immediately after* another concept, i.e. there are no other terms referencing a bounding box in between. For instance, for the first sentence in Figure 2, we consider *woman* to be followed by *dress*, *dress* followed by *boot*, and *boot* followed by *car*, but not *woman* followed by *car* or *boot*. Concept selection is performed in a greedy fashion, by choosing from the pool of bounding boxes for each image, the concept that is most likely to occur first in a sentence, followed by the concept that is most likely to occur given the previously selected concept. The selection process terminates when no remaining concept from the candidate pool is likely to follow the previously selected concept.



(a) Baselines based on textual priors



(b) Baselines based on visual cues

Figure 4: 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.

For all baselines, we select the first term among the synonyms of the WordNet synset to generate the referring expression for each concept.

4.3 Function Words

Our metric only evaluates the content selection process and ignores everything else. However, for completeness and in the spirit of generating complete descriptions, we attempt to connect selected concept terms with randomly selected function words or phrases. The phrases are selected to be a random word from a predefined list of prepositions and conjunctions, followed by an optional article *the*.

5 Experimental Results

The generation systems described in Section 4 were evaluated using the proposed content selection metric (Section 3.1). We also compared the proposed systems to a baseline that selects bounding boxes at random, up to a pre-defined threshold k of the maximum allowed number of bounding boxes per image. We explore different values of this threshold by varying k from 1 to 15. We take $min(k, N_{box})$ for images with fewer than k bounding boxes, where N_{box} is the total number of bounding boxes for the image.

As an upper bound to how well humans perform content selection, we evaluated the gold standard descriptions by evaluating one description against the other descriptions of the image and repeating the process for all descriptions. The upper bound is computed to be $F = 0.74 \pm 0.12$, with $P = 0.77 \pm 0.11$ and $R = 0.77 \pm 0.11$.

Figure 3 shows the F-scores of our proposed generation systems. Firstly, we examine the effects of varying the threshold k on the number of instances to be selected. The F-score peaks at k between 3 and 4 across all systems except the random baseline, and then drops or remains stagnant beyond that. Figure 4 gives an insight about this observation when the precision P and recall R are examined separately. As expected, P decreases while R increases when k is increased. The two graphs intersect at about k between 3 to 4, suggesting that these values are an optimal tradeoff between precision and recall (the mean number of unique instances per description is 2.89 in the development dataset).

Comparing the baselines based on visual and textual cues, the F-score in Figure 3 suggests that baselines using textual cues perform better when k is small, and visual cues perform better with larger k's. However, Figure 4 gives a clearer picture, where the bigram-based system obtained the best precision regardless of k (Figure 4a), while the systems based on bounding box cues relied on the increased recall when increasing k to obtain a high F-score (Figure 4b). Note that the bigram-based generation system is less sensitive to larger k's as the model itself contains an internal stopping criterion when no suitable concept is likely to follow a selected concept, resulting in a lower but stable recall rate compared to other systems, when k is increased. Figure 5 shows some example sentences generated by our baseline systems, for k=3.

We can infer from the results that (i) using prior

knowledge on the ordering of concepts (i.e. bigrams) is helpful for concept selection; (ii) frequency of concepts (i.e. unigrams) are helpful when there are only one or two instances to be described, possibly because the remaining objects are not mentioned as frequently as the main actors; (iii) visual cues are helpful for concept selection, although the precision is reduced as k increases.

5.1 Combining Textual and Visual Priors

We also explored combining textual priors and visual cues, which could potentially produce a stronger baseline. This is done by re-ranking the bounding boxes, for each image, by the average rank from both systems. In the case of the bigram-based system, bounding boxes that are not selected are all assigned an equal rank: $0.5 \times ((N_{box} + 1) - N_s) + N_s$, where N_{box} is the number of all bounding boxes for the image and N_s the number of bounding boxes selected by the bigram-based system. For example, if only 3 out of 9 bounding boxes are selected (and assigned ranks 1, 2 and 3 respectively), then the remaining 6 bounding boxes are all assigned equal rank 6.5. Figure 6 compares the F-scores of systems combining textual priors (unigram or bigram) and visual cues (bounding box position) at k=3 and k=4; we omitted bounding box size as the results are similar to bounding box position. Combining unigram and bounding box position did not significantly improve the F-score compared to using bounding box position alone, at k=3 and k=4. As seen earlier, the performance of the unigram-based system at these k's is much lower than systems based on visual cues. The combination of bigram and bounding box position, however, seems to yield slightly improved performance at these k's. This is likely due to the bigram-based system providing higher precision and the system based on visual cues providing better recall. This shows that combining textual and visual priors may be beneficial when they complement each other.

6 Discussion and Future Work

We presented the task of generating image descriptions from gold standard labelled bounding boxes as input to a text generation system. We also proposed a fine-grained evaluation metric specifically to evaluate the content selection capabilities of the image description generation system, which measures how well the system selects the concepts to be described compared against a set of human-authored reference descriptions. Several baselines were proposed to demonstrate the proposed metric on the task. We found that selecting a maximum of 3 to 4 instances is optimal for this dataset, and that both text and visual cues play a part in the content selection process.

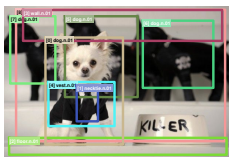
Further challenges can be observed for the proposed generation task based solely on gold standard visual inputs:

Bounding boxes. Bounding boxes labelled with concepts may be a good starting point for a 'clean' input task, but may be somewhat uninformative as important visual information is discarded in the process that might prove useful for the generation process. A possible solution would be to enrich the bounding box inputs with more information, either as attributes (adjectives, verbs etc.) or directly using visual features. However, manually annotating such fine-grained information is an onerous task.

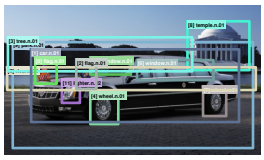
Suitability of metrics. Another possible issue with the proposed task is that it might be problematic to assume that all image description generation systems will be using a common pipeline. With the large variation in how image description generation systems are constructed, it may be difficult to constrain and expect systems to be using the same architecture that will enable us to evaluate them with such fine-grained metrics.

Future work with fine-grained metrics. Although we only consider one metric to evaluation the content selection capabilities of generation systems, further fine-grained metrics can be introduced to evaluate different components of the generation pipeline. Some examples include content ordering, lexicalisation or referring expression generation of concepts (and/or their attributes), evaluating the appropriateness of verbs, predicates and prepositions, and surface realisation.

Future work on image description generation. In this paper, we presented several baselines based on different textual and visual priors, and also explored combining cues from both text and vision. Future work on image description generation could involve stronger cues, for example from co-occurrences and spatial relationships between multiple objects.

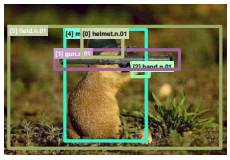


random:	[F=0.04]	[Wall]^{3} among [necktie]^{1} underneath [floor]^{2} .
bbox pos:	[F=0.00]	[Hallway]^{8} below the [wall]^{3} near the [floor]^{2} .
bbox size:	[F=0.39]	[Hallway]^{8} behind the [dog]^{0} underneath the [wall]^{3} .
unigram:	[F=0.05]	[Wall]^{3} near [floor]^{2} with the [dog]^{5} .
bigram:	[F=0.51]	[Dog]^{5} against [dog]^{0} beside the [dog]^{6} .

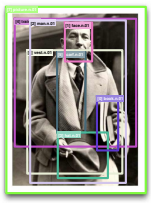


random:	[F=0.05]	[Park]^{9} behind [wheel]^{7} underneath the [window]^{6} .
bbox pos:	[F=0.59]	[Park]^{9} on the [car]^{1} below [river]^{5} .
bbox size:	[F=0.44]	[Park]^{9} behind the [car]^{1} against the [tree]^{3} .
unigram:	[F=0.42]	[Tree]^{3} beneath [car]^{1} by [window]^{6} .
bigram:	[F=0.71]	[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 [floor]^{1} below [wall]^{2} .
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=0.49]	[Picture]^{7} on [man]^{2} beside the [train]^{4} .
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: Example image descriptions generated by our baselines ($k = 3$).

We believe that the introduction of a fine-grained approach to evaluating image description generation tasks can encourage further growth in this area, linking further research between computer vision and natural language generation.

Acknowledgments

This work has been supported by the EU CHIST-ERA D2K 2011 Visual Sense project, EPSRC grant reference: EP/K019082/1.

References

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, pages 97–112.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ali Farhadi, Mohsen Hejrati, Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences for images. In *Proceedings of the European Conference on Computer Vision*.

Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas, and Krystian Mikolajczyk. 2015. Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In *CLEF2015 Working Notes*, CEUR Workshop Proceedings, Toulouse, France, September 8-11. CEUR-WS.org.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Ar-*

		P	R	F
Upper bound		0.77 ± 0.11	0.77 ± 0.11	0.74 ± 0.12
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	$k=1$	0.57 ± 0.41	0.21 ± 0.19	0.30 ± 0.25
	$k=2$	0.56 ± 0.27	0.40 ± 0.25	0.45 ± 0.25
	$k=3$	0.53 ± 0.20	0.54 ± 0.26	0.52 ± 0.21
	$k=4$	0.49 ± 0.16	0.65 ± 0.24	0.55 ± 0.17
	$k=5$	0.46 ± 0.14	0.72 ± 0.23	0.54 ± 0.15
	$k=6$	0.42 ± 0.13	0.77 ± 0.21	0.53 ± 0.14
	$k=7$	0.39 ± 0.13	0.81 ± 0.19	0.51 ± 0.12
	$k=8$	0.36 ± 0.12	0.84 ± 0.18	0.49 ± 0.12
	$k=9$	0.34 ± 0.12	0.87 ± 0.16	0.47 ± 0.12
	$k=10$	0.32 ± 0.12	0.89 ± 0.15	0.46 ± 0.12
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: 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).

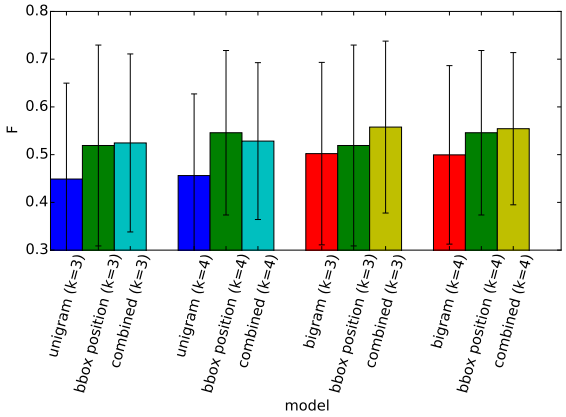


Figure 6: The content selection score, F, when combining textual priors and visual cues. For text priors, we compare both unigram and bigram priors. For visual cues, we show only the results for bounding box position as using bounding box size yields similar results. We compare the combined baselines at k=3 and k=4.

tificial Intelligence Research (JAIR), 47(1):853–899, May.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, Jeju Island, Korea, July. Association for Computational Linguistics.

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2013. Generalizing image captions for image-text parallel cor-

pus. In *The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Computational Natural Language Learning (CoNLL)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Tomas Mikolov, Martin Karafit, Luks Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 1045–1048. ISCA.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France, April. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*, pages 512–519.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2014. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575.

Richard Socher, Andrej Karpathy, Quoc Le, Christopher Manning, and Andrew Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Karl Stratos, Aneesh Sood, Alyssa Mensch, Xufeng Han, Margaret Mitchell, Kota Yamaguchi, Jesse Dodge, Amit Goyal, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. 2012. Understanding and predicting importance in images. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Mauricio Villegas, Henning Müller, Andrew Gilbert, Luca Piras, Josiah Wang, Krystian Mikolajczyk, Alba García Seco de Herrera, Stefano Bromuri, M. Ashraful Amin, Mahmood Kazi Mohammed, Burak Acar, Suzan Uskudarli, Neda B. Marvasti, José F. Aldana, and María del Mar Roldán García. 2015. General Overview of ImageCLEF at the CLEF2015 Labs. Lecture Notes in Computer Science. Springer International Publishing.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454. Association for Computational Linguistics.

Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song Chun Zhu. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 110–120, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

C. Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

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.

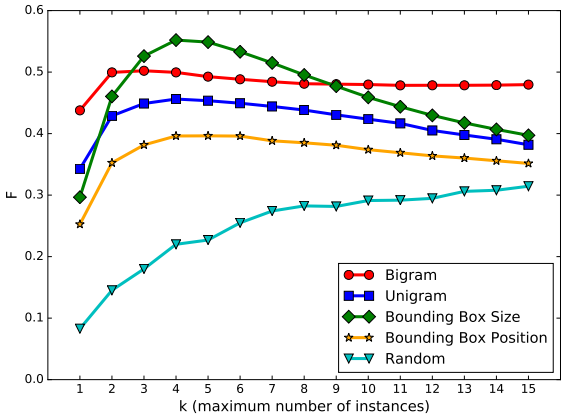


Figure 1: Replaces **Figure 3** of the original paper. The figure shows the content selection score, F, evaluated on the proposed baselines at varying levels of k (maximum number of instances per sentence). Standard deviations are omitted for clarity, but are included in Table 1.

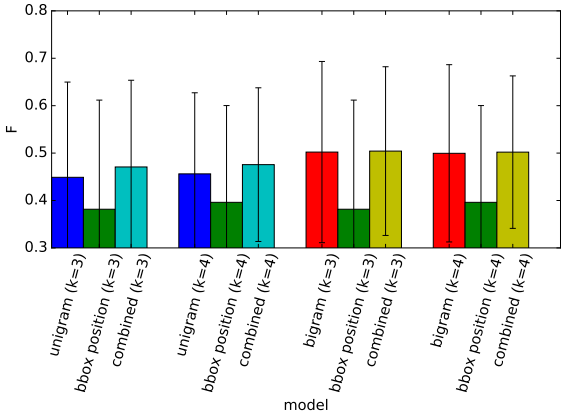


Figure 2: Replaces **Figure 6** of the original paper. The figure shows the content selection score, F, when combining textual priors (unigram or birgram) and visual cues based on **bounding box positions**. We compare the combined baselines at k=3 and k=4.

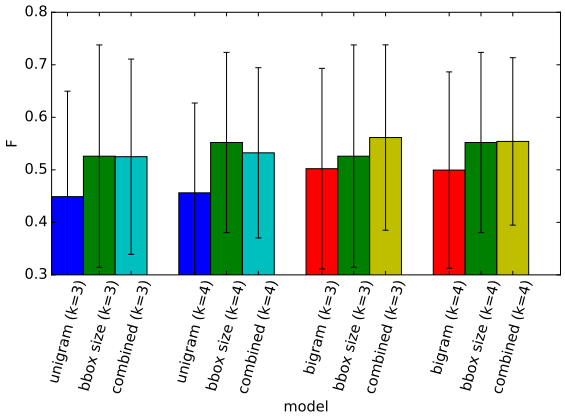
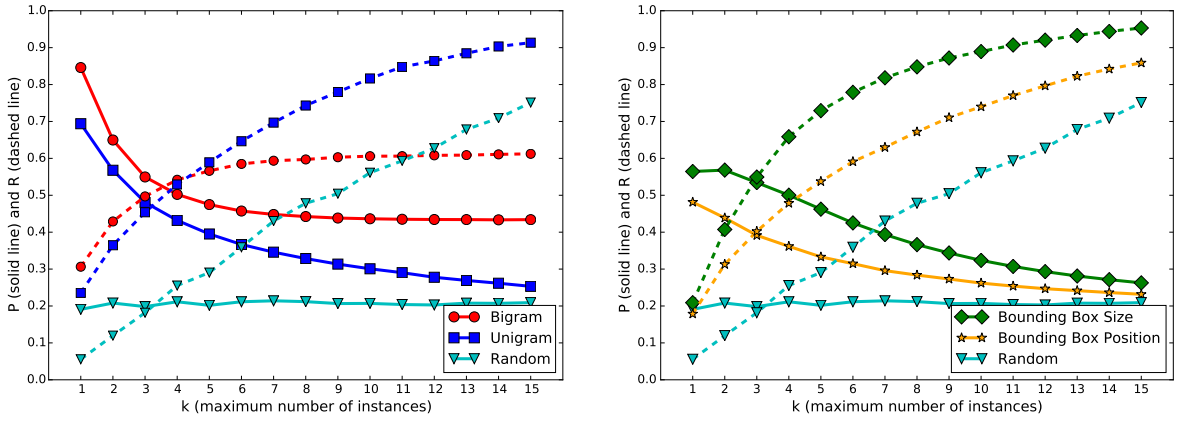


Figure 3: The content selection score, F, when combining textual priors (unigram or birgram) and visual cues based on **bounding box sizes**. We compare the combined baselines at k=3 and k=4. This figure is provided as a supplement, as our initial claim that using bounding box position and using bounding 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

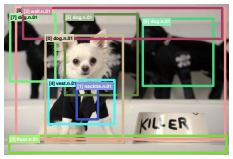
Table 1: Replaces **Table 1** of the original paper. The table shows the P, R and F scores (with standard deviations) of the content selection metric, as evaluated on different baselines at varying levels of k (1 to 10).



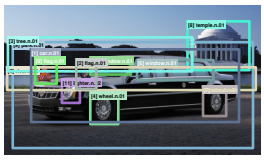
(a) Baselines based on textual priors

(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.

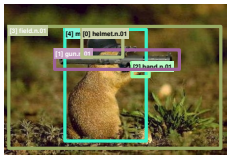


random:	[F=0.04]	[Wall]^{3} among [necktie]^{1} underneath [floor]^{2} .
bbox pos:	[F=0.45]	[Hallway]^{8} below the [dog]^{5} near the [dog]^{0} .
bbox size:	[F=0.39]	[Hallway]^{8} behind the [dog]^{0} underneath the [wall]^{3} .
unigram:	[F=0.05]	[Wall]^{3} near [floor]^{2} with the [dog]^{5} .
bigram: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

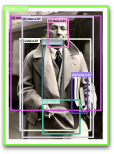


random:	[F=0.05]	[Park]^{9} behind [wheel]^{7} underneath the [window]^{6} .
bbox pos:	[F=0.59]	[River]^{5} on the [car]^{1} below [park]^{9} .
bbox size:	[F=0.44]	[Park]^{9} behind the [car]^{1} against the [tree]^{3} .
unigram:	[F=0.42]	[Tree]^{3} beneath [car]^{1} by [window]^{6} .
bigram:	[F=0.71]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