Binary Image Selection (BISON): Interpretable Evaluation of Visual Grounding

Hexiang Hu*
University of Southern California
hexiangh@usc.edu

Ishan Misra
Facebook AI Research
imisra@fb.com

Laurens van der Maaten
Facebook AI Research
lvdmaaten@fb.com

Abstract

Providing systems the ability to relate linguistic and visual content is one of the hallmarks of computer vision. Tasks such as image captioning and retrieval were designed to test this ability, but come with complex evaluation measures that gauge various other abilities and biases simultaneously. This paper presents an alternative evaluation task for visual-grounding systems: given a caption the system is asked to select the image that best matches the caption from a pair of semantically similar images. The system's accuracy on this Binary Image SelectiON (BISON) task is not only interpretable, but also measures the ability to relate fine-grained text content in the caption to visual content in the images. We gathered a BISON dataset that complements the COCO Captions dataset and used this dataset in auxiliary evaluations of captioning and caption-based retrieval systems. While captioning measures suggest visual-grounding systems outperform humans, BISON shows that these systems are still far away from human performance.

1. Introduction

Understanding and communicating visual content is a fundamental goal of intelligent agents. This goal has motivated a large body of research into systems that relate visual and linguistic information, *i.e.*, that perform *visual grounding*. Image captioning [23, 26, 55, 57] is a popular task to (holistically) test visual grounding by requiring systems to describe an image in natural language. Unfortunately, the open-ended nature of this task makes it difficult to develop "good" evaluation measures for captioning. In particular, common evaluation measures for captioning [3, 54] gauge more abilities than just visual grounding, such as fluency in language generation [1]. Moreover, the measures are difficult to interpret and appear to overestimate the performance of captioning systems by suggesting these systems exhibit super-human performance. As an alternative, some studies have evaluated caption-based image retrieval [23, 34]; however, retrieval measures make strong assumptions about (the lack of) negative data and are computationally intensive.

Figure 1: **Binary Image SelectiON (BISON):** Given a *target caption*, the system must select which of two images best matches the caption. This task evaluates fine-grained visual grounding. The BISON accuracy of a system is the proportion of examples for which the system correctly chooses the *target image* (✓) over the *decoy image* (✗).

This paper proposes an alternative task to evaluate the ability of a system to perform fine-grained visual grounding, called *Binary Image SelectiON (BISON)*. In BISON, the system is provided with two semantically similar images and a fine-grained text description that describes one image but not the other. The system needs to select which of the two images best matches the text description; see Figure 1. The performance of the system is measured in terms of its binary classification accuracy of selecting the correct image, which is an interpretable measure. BISON can be used as an auxiliary evaluation of generative (captioning) and discriminative (retrieval) vision-language models[1],

*This work was performed while Hexiang Hu was at Facebook.

[1]We note that BISON is by no means an evaluation that captures all aspects that are important in vision-language models. For instance, it does not assess language generation quality.

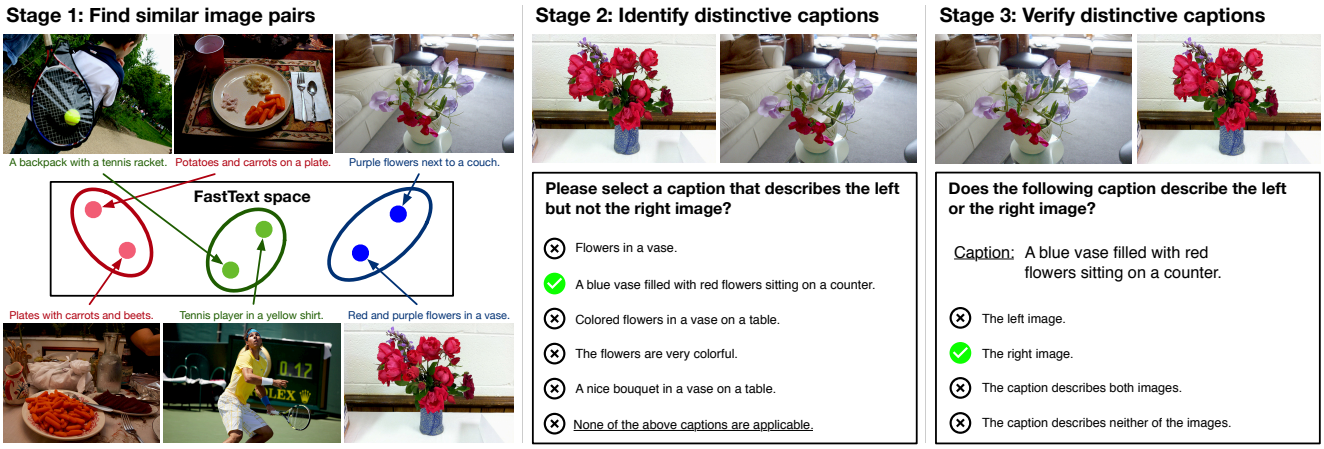


Figure 2: **Illustration of COCO-BISON dataset collection:** We collect the dataset for our binary image selection task using the COCO captions dataset. We first find similar images using description-based similarity. Annotators then select a caption that describes only one of the images in a pair. Finally, we validate the annotation by asking separate annotators to pick correct the image using the description. See Section 4 for details.

which facilitates its use in conjunction with existing evaluations. BISON accuracy differs from existing captioning evaluations in that it focuses on fine-grained visual content, and from retrieval evaluations in that it provides guaranteed negative examples rather than relying on noisy labelings in which "negative" examples may, in fact, be positive.

To facilitate binary image selection experiments, we collected the *COCO-BISON Dataset* using the images and captions in the existing COCO [12] validation set. By using both the text and images from the COCO dataset, we ensure that COCO-BISON has a similar distribution as COCO — this allows for the evaluation of COCO-trained models on the COCO-BISON dataset. Because we believe that any good model that relates vision and language should be able to perform fine-grained visual grounding, we use COCO-BISON only to evaluate (and not train) existing methods. We use the COCO-BISON dataset to evaluate the differences between BISON and alternative tasks, and provide a detailed BISON benchmark of state-of-the-art captioning and caption-based retrieval systems.

2. Related Work

There exists a large body of prior work on relating visual and linguistic information, focusing on a variety of different tasks and experimental setups. Primary examples of tasks that involve visual grounding include referring expressions [27, 30, 59], visual story-telling [24], visual question answering [7, 38], visual question generation [16, 39, 42], and zero-shot learning [2, 47]. A comprehensive overview of all this prior work is outside the scope of this paper; we refer the reader to [19] for a survey. Most of these tasks are more AI-complete than binary image selection in the sense that they simultaneously assess a range of system abilities that go beyond assessing just visual grounding.

Binary image selection is closely related to image captioning, image retrieval, and referring expression tasks. We provide a brief overview of work in these domains below.

Image captioning [6, 13, 18, 23, 26, 32, 55, 57] is a task in which the system generates a textual description of an image. The task assesses a system's ability to ingest visual information in an image and generate linguistically fluent natural language descriptions of that information [12]. As a result, image captioning gauges not only visual understanding but also the generative linguistic prowess of systems. In contrast, binary image selection only evaluates the ability of a system to discriminate between images based on a text description — without conflating it with generation.

A recent line of work [5, 36, 51, 53] focuses on generating discriminative text descriptions for images. BISON is related to this work but it focuses exclusively on the discriminative aspect of the task by not considering language generation. As a result, BISON gives rise to a more stable and interpretable evaluation measure of visual grounding.

In image captioning, human evaluation of the generated captions is generally considered the gold standard. As human evaluation is cumbersome, many automatic evaluation measures [3, 8, 35, 43, 54] have been proposed that ease the evaluation of captioning systems. These measures compute a similarity between the generated captions and a set reference captions. Because humans tend to describe images differently, many such reference captions must be collected to obtain a robust evaluation [11, 40]. However, captioning datasets [12, 58] only contain a handful of reference captions for each image, which impacts the quality of the evaluation measures (see Section 5). We discuss and benchmark

a range of image captioning systems in Section 6.

Image retrieval [9, 10, 20, 23, 44, 50] is a task in which the system is asked to retrieve relevant images given a text description (or vice versa). Retrieval performance is generally measured in terms of recall@k [23]. Similar to BISON, caption-based image retrieval evaluates how well a system can distinguish relevant images from irrelevant ones. The key difference between image retrieval and BISON is that retrieval evaluations rely on "implicit" negatives: retrieval datasets provide manually annotated positive image-description pairs, but they *assume* that every image-description pair that is not in the dataset is a negative example. In practice, this assumption is often violated: many such image-description pairs would actually be labeled positively by a human annotator [34]. In contrast to retrieval datasets, each example in our COCO-BISON dataset contains a positive and a *genuinely* negative image-description pair, which facilitates more reliable evaluation. We analyze the sensitivity of retrieval measures in Section 5.2 and benchmark image retrieval systems in Section 6.

Referring expressions [15, 27, 30, 41, 59] is a task in which a system is asked to distinguish objects in a *single* image based on a text description. BISON can thus be viewed as a kind of *holistic* referring-expressions task that involves between-image rather than within-image comparisons. In contrast to referring expressions that focus on a single object and its attributes, text descriptions in BISON may focus on groups of objects and their attributes, relationships between these objects, and even entire scenes.

3. Analysis of Existing Captioning Evaluations

In contrast to BISON accuracy, automatic captioning evaluations such as BLEU-4 [43], CIDEr [54], METEOR [8], and SPICE [3] compare a generated caption to a collection of reference captions. As a result, the evaluations may incorrectly assess the semantics of the generated caption and they may be sensitive to changes in the reference caption set. We perform two experiments designed to study these effects. We perform both experiments on the COCO validation set, using the state-of-the-art [49] UpDown [4] captioning system[2] to generate the captions.

Semantic quality of captioning evaluations. Automatic captioning evaluations may incorrectly assess the quality of a generated caption because they may fail recognize that a generated caption is semantically similar to the reference captions because it uses a different phrasing, or because the generated caption may focus on aspects of the visual content that are not described in the reference captions. To gauge the extent to which these issues hamper automatic evaluations of captioning systems, we asked human annotators to evaluate the semantic quality of generated captions; we compare the resulting quality evaluations with those obtained via four automatic captioning evaluations.

To obtain human annotations, we followed the COCO guidelines for human evaluation [1] and asked annotators to evaluate the "correctness" of image-caption pairs on a Likert scale from 1 (low) to 5 (high). We asked a second set of annotators to evaluate the "detailedness" of captions (without showing them the image) on the same Likert scale.

Figure 5 shows the resulting correctness and detailedness assessment as a function of four automatic captioning evaluations that were normalized to lie between 0 and 1. The results in the figure suggest that automatic captioning evaluations are not very predictive of the correctness of generated captions, and do not encourage these captions to be very detailed. Figure 3 shows three qualitative examples of generated captions with low CIDEr score but a high correctness score. Together, these results highlight the limitations of using a handful of reference captions to evaluate captioning systems: the reference captions do not capture all visual content and all the different ways in which that content can be described [11, 40]. This leads captioning measures to reward systems for generating very generic captions.

Effect of the number of reference captions. To assess the robustness of automatic captioning evaluations, we measured captioning scores using reference-caption sets of varying size. To construct these reference-caption sets, we selected captions uniformly at random (without replacement) from the five reference captions provided for each image in the COCO captions dataset.

Figure 4 shows the value of each captioning score relative to the value of that score evaluated using all five reference captions (right-most point). We show this value as a function of the cardinality of the reference caption subset. The results presented in the figure show that three out of four captioning measures are sensitive to the number of reference captions that the caption dataset provides. BLEU-4 appears to be the most sensitive measure: whilst using all five reference captions leads to a BLEU-4 score of 34.58, using a single reference caption reduces the BLEU-4 score to just 10.69 (see supplementary material for other measures). In contrast to all other evaluation measures, the SPICE score decreases with more reference captions.

4. The COCO-BISON Dataset

Motivated by the analysis of captioning measures above, we develop binary image selection (BISON) with the aim of providing a robust, auxiliary evaluation of visual grounding that rewards systems for generating detailed, discriminative captions. To do so, we collect BISON annotations on top of validation split of the COCO captions dataset [12].

[2] We performed the same experiment using other captioning methods. The results of these experiments were qualitatively similar, and are presented in the supplemental material.

Figure 3: **Correct image-caption pairs:** All pairs have a correctness score of 4.0 (as rated by human annotators), but a low CIDEr score. The low CIDEr score can be attributed to having few reference captions to evaluate against.

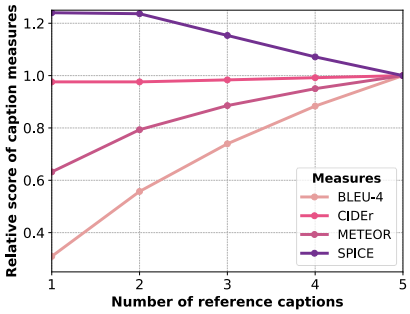


Figure 4: **Captioning scores vs. number of reference captions.** Captions generated using the UpDown [4] model. Scores are *relative* to their value when using all five reference captions.

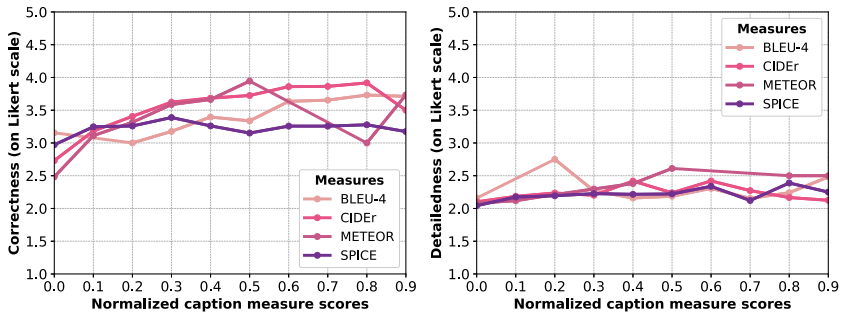


Figure 5: **Correctness** (left) **and detailedness** (right) **of generated captions as a function of their captioning scores.** Captions were generated using the UpDown [4] captioning system. Correctness and detailedness of the generated captions were rated on a Likert scale (from 1 to 5) by human annotators. The average correctness and detailedness scores are 3.266 and 2.203, respectively.

4.1. Collection of BISON Annotations

Figure 2 illustrates our collection of binary image selection annotations that comprises the following three stages.

1. Collect pairs of semantically similar images. We construct a semantic representation for each image in the COCO validation set by averaging word embeddings (obtained using FastText [25]) of all the words in all captions associated with the image. We use these representations to find the semantically most similar image for each "target" image in the dataset via nearest neighbor search. We refer to the nearest neighbor of a target image as the "decoy" image.

2. Identify captions that distinguish targets and decoys. We present human annotators with an interface[3] that shows: (1) a target image, (2) the corresponding decoy image, and (3) the five captions associated with the target image in the COCO captions dataset. We ask the annotators to select a caption from the set of five that describes the target image *but not* the decoy image, or to select "none of the above" if no discriminative caption exists. Unless annotators select the latter option, each of their annotations produces a caption-target-decoy triple. We discard all image pairs for which annotators indicated no discriminative caption exists.

3. Verify correctness of the caption-target-decoy triples. To ensure the validity of each caption-target-decoy triple, we presented a different set of human annotators with trials that contained the target and decoy images and the caption selected in stage 2. We asked the annotators whether the selected caption describes: (1) the target image, (2) the decoy image, (3) both images, or (4) neither of the images. Each verification trial was performed by two annotators; we only accepted the corresponding BISON example if both annotators correctly selected the target image given the caption.

The caption-target-decoy triples thus collected form binary image selection (BISON) examples, two of which are shown in Figure 1. The COCO-BISON dataset is available from http://hexiang-hu.github.io/bison.

[3] Screenshots of the annotation interface in the supplementary material.

	Flickr-30K	COCO val	COCO-BISON
Number of examples	5,070	202,654	54,253
Unique images	1,014	40,504	38,680
Unique captions	5,068	197,792	45,218

Table 1: **Key statistics of our COCO-BISON dataset** in comparison to the Flickr-30K [58] and COCO Captions [12] validation sets.

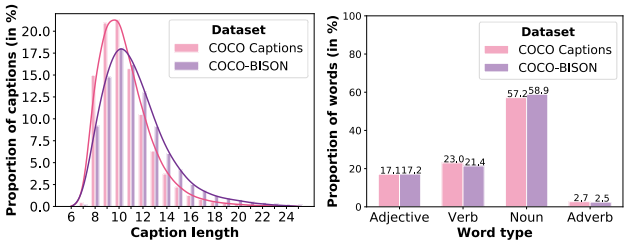


Figure 6: **Caption statistics of COCO and COCO-BISON:** Length distribution (left) and part-of-speech distribution (right) of captions in the datasets.

4.2. Dataset Characteristics

Table 1 presents key statistics of our COCO-BISON dataset, comparing them to the statistics of the validation splits of two popular captioning datasets. The statistics reveal that our three step annotation procedure (Section 4.1) identified a BISON example for $38,680/40,504 \approx 95.5\%$ of the images in the COCO validation set.

4.3. Definition of the BISON Task

In the BISON task, the model is provided two images and a sentence description that applies to only one of the images (see Figure 1). The model is then asked to pick the correct image and its performance can be measured using binary classification accuracy. We refer to this accuracy as the BISON score and report the mean accuracy over the COCO-BISON data. The BISON task is only used for evaluation and no methods are trained on this data.

Binary image selection facilitates evaluations of both image captioning and image retrieval systems. To evaluate both types of systems, we compute "compatibility" scores between the text description and each of the two images, and pick the image with the higher score. For image captioning systems, this score is defined as the log-likelihood of the text description given the image. The image retrieval systems naturally compute the compatibility score, *e.g.*, by using an inner product of the image and text features.

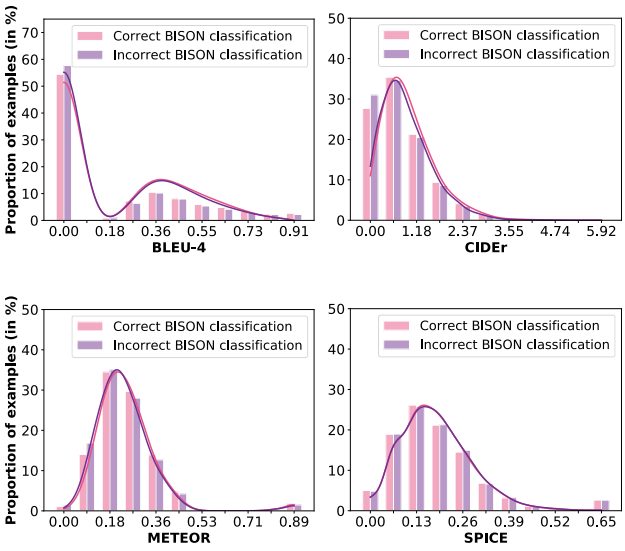


Figure 7: **Distribution of captioning scores** (BLEU-4, CIDEr, METEOR, and SPICE) for BISON examples that were classified incorrectly (pink bars) and correctly (purple bars). Captions generated using the UpDown [4] system.

5. Comparing BISON to Existing Evaluations

We performed a series of experiments that aim to study how evaluating systems via BISON compares to existing evaluations based on image captioning or caption-based image retrieval. We present the results of our experiments for captioning and caption-based retrieval separately below.

5.1. Comparing BISON and Captioning

We conduct two experiments aimed at comparing BISON accuracy to existing captioning evaluations.

Does BISON accuracy predict captioning scores? We evaluate the UpDown captioning system in terms of BISON accuracy and in terms of four captioning scores on the COCO-BISON dataset. Figure 7 shows the distribution of captioning scores for the target images from correctly and incorrectly classified BISON examples separately. Specifically, for each COCO-BISON example that the model classifies correctly (or incorrectly), we generated a caption for the target image and measured the captioning score of the generated caption. The figure shows that distribution of all the captioning scores is nearly identical for BISON examples that were correctly and incorrectly classified. This weak relation suggests that BISON assesses different aspects of visual grounding than captioning.

Do caption score differences provide signal for BISON? We try and classify BISON examples based on captioning scores for the target and decoy images in those examples. Specifically, we compute a captioning score (*e.g.*, BLEU-4)

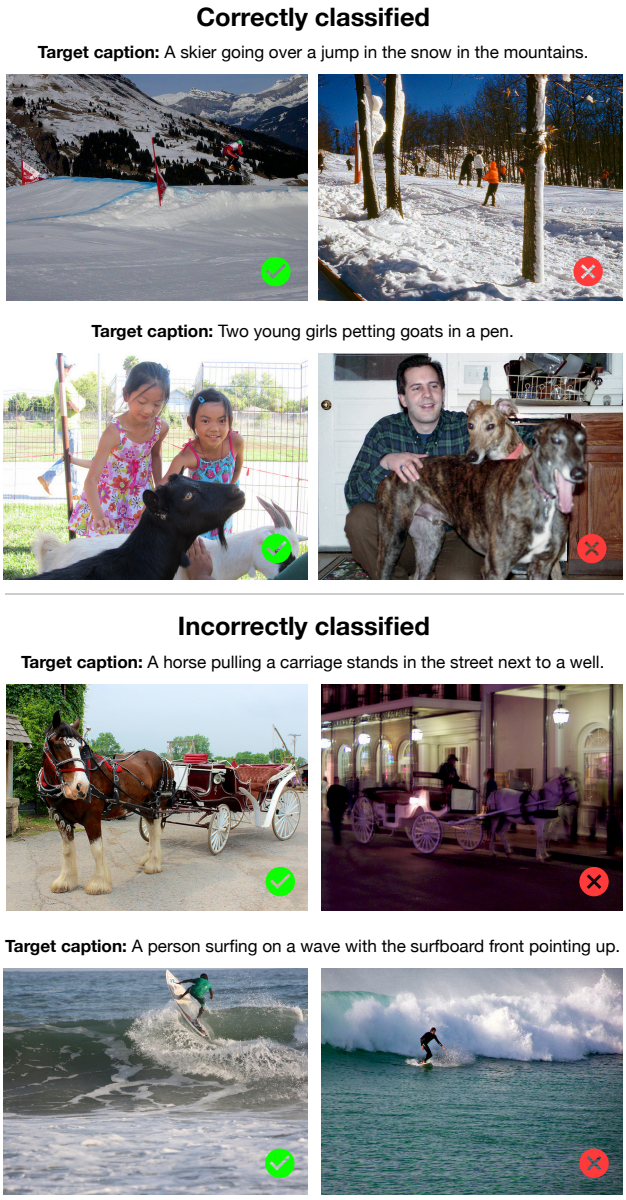


Figure 8: **Correctly and incorrectly classified BISON examples.** Classifications were performed by measuring the CIDEr score of the caption generated by the UpDown [4] captioning system for both images, and selecting the image with the highest score. Correct predictions are shown at the top and and incorrect predictions at the bottom.

between the target caption and the COCO reference captions for the target and decoy images. When computing the score for the target image, we remove the target caption from the reference captions; for the decoy image, we randomly select four captions from the reference captions (without replacement). Next, we select the higher-scoring image as prediction for the BISON example. Figure 8 shows examples of COCO-BISON samples that were classified correctly and incorrectly using this approach.

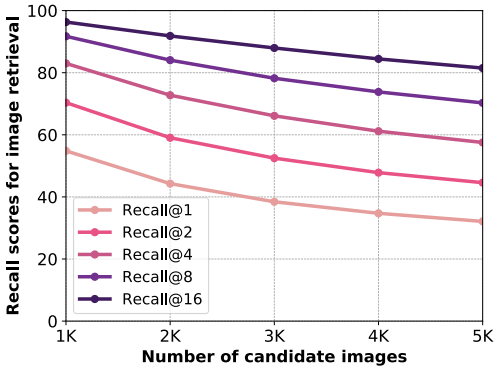


Figure 9: **Recall@k of image retrieval as a function of the size of the image dataset.** Caption-based image retrieval was performed using the SCAN [33] (t2i) retrieval system.

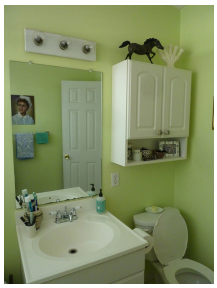
The BISON accuracy of this approach is 70.73% for BLEU-4, 70.78% for CIDEr, 74.44% for METEOR, and 62.79% for SPICE. This result shows us that predictions based on captioning scores incorrectly select the decoy image at least 25% of the times, *despite relying on access to the ground-truth* reference captions. This low prediction accuracy suggests that BISON accuracy and captioning scores are, indeed, different measures for visual grounding quality.

5.2. Comparing BISON and Image Retrieval

Another common method for assessing visual grounding systems is by using them for caption-based image retrieval or image-based caption retrieval. The performance on these tasks is measured via recall@k. Binary image selection (BISON) is closely related to caption-based image retrieval in that it requires the system to produce a ranking of images given a caption. There are two key differences between BISON and caption-based image retrieval.

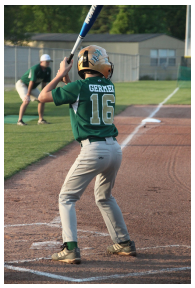
The first difference is that BISON involves ranking only two images rather than the entire image collection. As a result, evaluating a system's BISON score is much less computationally intensive: getting a BISON score for a single example requires scoring only two image-caption scores using the system, whereas a caption-based image retrieval evaluation would require scoring N image-caption pairs if the image collection contains N images. This makes computing the BISON score of a system more practical than computing its recall@k. Moreover, BISON scores are better calibrated: they do not change when the size of the image collection changes — this is in contrast to recall@k measures that are very sensitive to the values of N and k.

The second difference between BISON and recall@k is that the image that is supposed to rank lower is guaranteed to be a true negative example in BISON. By contrast, many



Caption query: A toilet sitting in a bathroom next to a sink.

Caption query: A cat standing on the toilet bowl seat.



Caption query: A young baseball player in uniform getting ready to bat.

Figure 10: **Examples of "incorrect" image retrievals for a caption query.** The examples suggest many "negative" examples in caption-based retrieval are actually positive.

of the "negative" images in caption-based image retrieval are actually positive examples for the query caption; see Figure 10 for examples. Because BISON does not average over incorrect labels, BISON scores are more reliable.

6. BISON Evaluation of State-of-the-Art Captioning and Retrieval Systems

We use the newly collected COCO-BISON dataset to evaluate many state-of-the-art captioning and caption-based image retrieval systems on the binary image selection task.

6.1. Evaluated Captioning and Retrieval Systems

We analyze three **image captioning systems**: (1) the *ShowTell* captioning system [55]; (2) an extension of the ShowTell system that can attend to specific parts of the image, called *ShowAttTell* [46, 57]; and (3) the state-of-the-art *UpDown* captioning system [4]. Like ShowAttTell, the UpDown system uses a spatial attention mechanism but it differs from ShowAttTell in that it uses two LSTMs: one for decoding captions and another one for generating spatial attention over image features.

We train all three captioning systems on the COCO Captions [12] training set by minimizing the cross-entropy loss per word over a vocabulary of $9,487$ words, and average the loss over all words in the caption. Following common practice in the literature [4, 36, 46], we also finetune the trained systems using self-critical sequence training (SCST; [46]). SCST uses the REINFORCE algorithm [52] to directly maximize the CIDEr score [54] of the captioning system. For completeness, we report the performance of the captioning systems both before and after SCST finetuning.

We also analyze four systems for **caption-based image retrieval**: (1) ConvNet+BoW, (2) ConvNet+Bi-GRU, (3) Obj+Bi-GRU, and (4) SCAN [33]. The *ConvNet+BoW* system represents the caption by averaging word embeddings over all words in the caption, and represents the image by averaging features produced by a convolutional network over regions (described later). The resulting representations are processed separately by two multilayer perceptrons (MLPs). We use the cosine similarity between the outputs of the two MLPs as an image-caption compatibility score. The *ConvNet+Bi-GRU* system is identical to the previous system, but it follows [29] and uses a bi-directional GRU [14] to construct a caption representation. The *Obj+Bi-GRU* system is similar to ConvNet+Bi-GRU but uses a Bi-GRU to aggregate image-region features (spatial ConvNet features or object proposal features) and construct the image representation. Finally, *SCAN* [33] is a state-of-the-art image-text matching system based on image-region features and stacked cross-attention; we implement two variants of this system, *viz.* one that uses image-to-text (i2t) attention and one that uses text-to-image (t2i) attention. All caption retrieval systems are trained to minimize a max-margin loss [17].

Dataset →	COCO validation split				COCO-BISON
Measure →	**BLEU-4**	**CIDEr**	**SPICE**	**METEOR**	**BISON**
Cross-entropy loss					
ShowTell [55]	32.35	97.20	18.34	25.51	78.59
ShowAttTell [57]	33.49	101.55	19.16	26.06	82.04
UpDown [4]	34.53	105.40	19.86	26.69	84.04
Self-critical sequence loss [46]					
ShowTell [55]	32.38	97.88	18.42	25.68	78.79
ShowAttTell [57]	33.99	103.68	19.53	26.37	82.73
UpDown [4]	**34.58**	**106.30**	**20.01**	**26.92**	84.27
Human [1]	21.7⋆	85.4⋆	19.8⋆	25.2⋆	**100.00**

Table 2: **Performance of three image captioning systems** in terms of four captioning scores on the COCO validation set (left) and in terms of BISON accuracy on the COCO-BISON dataset (right). Human performances marked with ⋆ were measured on the COCO test set. See text for details.

Implementation Details. Following the current state-of-the-art in image captioning [4, 33], all our systems use the top 36 object proposal features produced by a Faster R-CNN model [45] with a ResNet-101 backbone that was trained on the ImageNet [48] and Visual Genome [31] datasets. In all models, word embeddings were initialized randomly. We refer the reader to the supplementary material for a complete overview of the hyper-parameters we used to train our models.

6.2. Results

Table 5 presents the BISON accuracy of our three **image captioning** systems on the COCO-BISON dataset. For reference, the table also presents the performance of these systems in terms of four standard captioning scores on the standard COCO validation set, and the performance of human annotators on the COCO test set (adopted from [1]).

Dataset →	COCO-1K [26]				COCO-BISON
Task →	**Image retrieval**		**Caption retrieval**		
Measure →	**R@1**	**R@5**	**R@1**	**R@5**	**BISON**
ConvNet+BoW	45.19	79.26	56.60	85.70	80.48
ConvNet+Bi-GRU [29]	49.34	82.22	61.16	89.02	81.75
Obj+Bi-GRU	53.97	85.26	66.86	91.40	83.90
SCAN i2t [33]	52.35	84.44	67.00	92.62	84.94
SCAN t2i [33]	**54.10**	**85.58**	**67.50**	**92.98**	**85.89**

Table 3: **Performance of five caption-based image retrieval systems** using recall@k ($k = 1$ and $k = 5$) on caption-based image retrieval and image-based caption retrieval on the COCO-1K dataset (left) and in terms of BISON accuracy on the COCO-BISON dataset (right).

The results reveal that, even though BISON measures different aspects of a system than the other captioning scores, the ranking of the three systems is identical across all evaluation scores. In line with prior work [49], we find that the UpDown captioning system outperforms its competitors in terms of all evaluation measures, including BISON.

The main difference between BISON and existing captioning scores is in how they rank the ability of humans to generate captions: all three systems outperform humans in terms of nearly all captioning scores, but they all perform substantially worse than humans in terms of BISON accuracy[4]. Unless one believes that current image captioning systems actually exhibit super-human performance (we do not), this suggests that measuring the BISON score of a system provides a more realistic assessment of the capabilities of modern image captioning systems compared to humans.

Table 6 presents the BISON accuracy of five **caption-based retrieval systems** on the COCO-BISON dataset. For reference, the table also presents the recall@k (for $k=1$ and $k=5$) of these systems on a caption-based image retrieval and an image-based caption retrieval task; these results were obtained on the COCO-1K split of [26]. As in our results for captioning systems, we observe that the ranking of caption-based retrieval systems in terms of BISON accuracy is identical to their ranking in terms of retrieval measures. In line with prior work [33], we find that the SCAN system with text-to-image (t2i) attention outperform the competing systems in terms of all measures.

7. Discussion

This study has explored binary image selection (BISON) as an alternative experimental setup for gauging the performance of systems that relate visual and linguistic content. Our empirical evaluations in the BISON paradigm revealed that binary image selection assesses a different set of capabilities than image captioning tasks, in particular, by focusing on "fine-grained" information in the language rather than the "generic" descriptions common in captioning. In a sense, BISON can be viewed as a variant of referring-expressions tasks that considers images "holistically" rather than focusing on image parts. BISON is closely related to the caption-based image retrieval task but has the advantage that the evaluation is more reliable, easily interpretable, and that it provides a calibrated score. Having said that, the BISON paradigm also has disadvantages compared to the tasks such as image captioning: for instance, it does not assess the fluency of generated captions. Therefore, we view binary image selection as an evaluation task that ought to be used *in conjunction* with other caption-related evaluations.

We observed that the relative ranking of modern systems in terms of captioning or retrieval scores is nearly identical to the ranking of those systems in terms of BISON. A potential explanation for this observation may be that some of the systems are simply unequivocally better than others: for instance, if system A has an image-recognition component that is substantially better than the image-recognition component of system B, it is quite likely that system A will outperform system B in a very wide range of tasks involving vision and language. We do like to emphasize, however, that it is well possible that the observed rank correlation between captioning and BISON scores may no longer hold when researchers start designing systems with the BISON evaluation in mind. Results comparing the performance of humans with that of our systems underline this point: existing captioning scores suggest that systems possess super-human capabilities, which contradicts human assessments of the quality of captions generated by the systems. By contrast, the BISON scores of current systems appear to be better aligned with human assessments of caption quality.

To conclude, we hope that the binary image selection task will foster research into models that go beyond coarse-level matching of visual and linguistic content by rewarding systems that can perform visual grounding at a detailed level. The interpretability of BISON makes it easier to debug and analyze this visual grounding. We hope that the public release of the COCO-BISON dataset will help the community assess whether we are making progress towards the goal of developing such systems.

[4]Please note that the accuracy of humans on the BISON task is 100% by definition due to the way the COCO-BISON dataset was collected.

Acknowledgements

We thank Devi Parikh, Marcus Rohrbach, and Brian Knott for comments on early versions of this paper.

References

[1] Microsoft COCO 1st Captioning Challenge (Large-scale Scene UNderstanding Workshop, CVPR 2015). http://lsun.cs.princeton.edu/slides/caption_open.pdf. Accessed: Nov 3, 2018. 1, 3, 7, 12, 13, 15