Evaluating Text-to-Image Matching using Binary Image Selection (BISON)

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 text-based image retrieval and image captioning were designed to test this ability, but come with evaluation measures that have high variance or are difficult to interpret. We study an alternative task for systems that match text and images: given a text query, the system is asked to select the image that best matches the query from a pair of semantically similar images. The system's accuracy on this Binary Image SelectiON (BISON) task provides a robust and interpretable measure of its ability to match linguistic content with fine-grained visual structure. We gather a BISON dataset that complements the COCO dataset and use it to evaluate modern text-based image retrieval systems.

1. Introduction

Understanding the relation between linguistic and visual content is a fundamental goal of computer vision, motivating a large body of research on tasks such as image retrieval and captioning. These tasks have challenges in terms of evaluation: in particular, the open-ended nature of image captioning tasks makes it difficult to develop reliable evaluation measures [1, 14], and text-based image retrieval is unreliable because retrieval datasets are only partly labeled: they incorrectly assume that images that are not positively labeled for a given text query are negative samples. Motivated by these issues, we propose an alternative task to evaluate systems that match textual and visual content, called *Binary Image SelectiON (BISON)*. In BISON, the system is provided with two similar images and a fine-grained text query that describes one image but not the other. The system needs to select which of the two images is described in the text query; see Figure 1. The performance of the system is measured in terms of its binary classification accuracy of selecting the correct image. BISON can be viewed as a variant of text-based image retrieval in which positive and negative examples are *explicitly labeled*, and can be used to evaluate both generative and discriminative vision-language models. BISON accuracy differs from existing tasks in that it is reliable, easily interpretable, and focuses on fine-grained visual content. To facilitate BISON experiments, we collected the *COCO-BISON* dataset using the images and captions in the existing COCO [4] validation set. We use the COCO-BISON dataset to evaluate state-of-the-art text-based image retrieval systems, shedding new light on the performance of these systems.

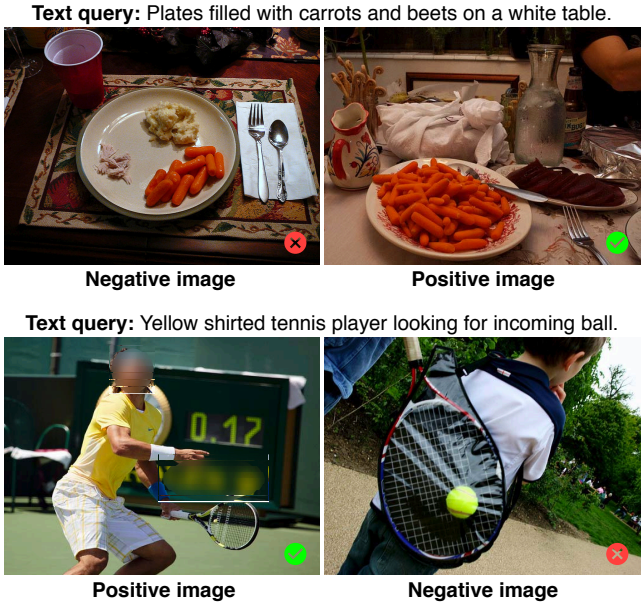

Figure 1: **Binary Image SelectiON (BISON):** Given a *text query*, the system must select which of two images best matches the caption. The BISON accuracy of a system is the proportion of examples for which the system correctly chooses the *positive image* (✓) over the *negative image* (✗).

2. Analyzing Retrieval and Captioning Tasks

We performed two experiments to identify the limitations of popular evaluations of vision-and-language systems via text-based image retrieval and image captioning.

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

Recall@1	Human	Number	Percentage
Incorrect	Incorrect	165	43.9%
Incorrect	Correct	211	56.1%

Table 1: **Analysis of the recall@1 text-based image retrieval measure.** We run SCAN t2i [10] image retrieval on the COCO captions [4] validation set and ask humans to analyze all 376 retrieval "errors" according to recall@1.

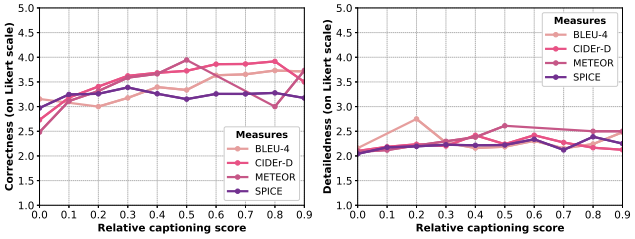

Figure 2: **Correctness** (left) **and detailedness** (right) **of generated captions as a function of their captioning scores.** Captions were generated using the UpDown [2] captioning system. Correctness and detailedness of the generated captions were rated on a Likert scale (from 1 to 5) by human annotators. The average correctness and detailedness scores are 3.266 and 2.203, respectively.

Text-based image retrieval. Evaluations via text-based image retrieval use a single positive image for each text query and assume all other images in the dataset are negative examples for that query [6]. This assumption is often incorrect, in particular, when the image datasets is large. To assess the severity this problem, we performed an experiment in which we analyzed the "errors" made by the state-of-the-art SCAN t2i retrieval system [10] on the COCO captions validation set. We presented each incorrectly retrieved image along with the text query to a set of human annotators, asking them to indicate if the text query appropriately describes the content of the image. The results of this experiment are presented in Table 1 and suggest that more than half of the "errors" made by the SCAN t2i system are not errors: the retrieved images are erroneously marked as incorrect due to the lack of explicit negative annotations.

Image captioning. Captioning evaluation measures such as BLEU-4 [11], CIDEr-D [14], METEOR [3], and SPICE [1] compare a generated caption to a collection of reference captions. As a result, the evaluations may be sensitive to changes in the reference caption set and incorrectly assess the semantics of the generated caption. We perform an analysis designed to study these effects on the COCO captions validation set by asking human annotators to assess image captions generated by the state-of-the-art UpDown [2, 13] captioning system. Specifically, we followed the COCO guidelines for human evaluation 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 2 shows the resulting correctness and detailedness assessments as a function of four captioning scores (BLEU-4, CIDEr-D, METEOR, and SPICE) that were normalized to lie between 0 and 1. The results in the figure suggest that captioning scores do not correlate with the correctness of generated captions very well, and do not encourage generated captions to provide a detailed description of the image.

3. The COCO-BISON Dataset

We collected binary image selection annotations for the validation split of the COCO captions dataset [4].

3.1. Collection of BISON Annotations

Figure 3 illustrates the three main stages of our pipeline for collecting binary image selection annotations.

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 [7]) of all the words in all captions associated with the image. We use these representations to find the semantically most similar image for each image in the dataset via nearest neighbor search. We label the query image as positive and its nearest neighbor as negative.

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

3. Verify correctness of the query-positive-negative triples. To ensure the validity of each query-positive-negative triple, we presented a different set of human annotators with trials that contained the positive and negative images and the query selected in stage 2. We asked the annotators whether the text query describes[1]: (1) the positive image, (2) the negative 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 positive image given the text query.

[1]In the verification stage, the annotators do not know which image is positive and which one is negative.

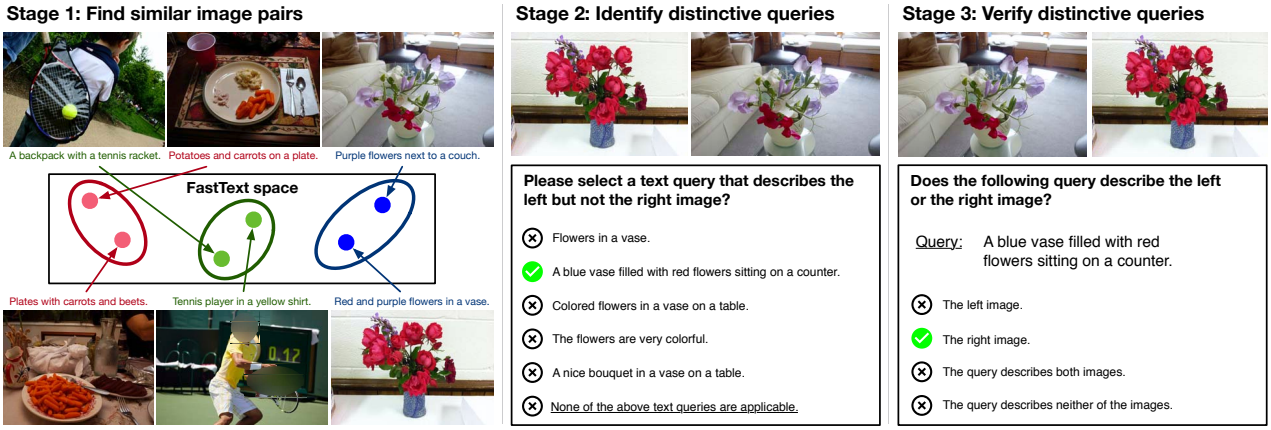

Figure 3: **Illustration of COCO-BISON dataset collection:** We collect annotations for our binary image selection task on top of the COCO Captions dataset. We first find pairs of semantically similar images based on the similarities between their reference captions. Annotators then select a text query that describes only one of the images in a pair. Finally, we validate the annotations by asking annotators to select the correct image given the text query. See Section 3 for details.

	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 2: **Key statistics of the COCO-BISON dataset:** The statistics of the Flickr-30K [15] and COCO Captions [4] validation sets are shown for reference.

The query-positive-negative triples thus collected form binary image selection (BISON) examples, two of which are shown in Figure 1. The COCO-BISON dataset and corresponding evaluation code is publicly available from http://hexiang-hu.github.io/bison/.

3.2. Dataset Characteristics

Table 2 presents key statistics of our COCO-BISON dataset; for reference, it also shows the statistics of the validation splits of two popular captioning datasets. As shown in the table, our three-stage annotation procedure produced a BISON example for $38,680$ of the $40,504$ the images in the COCO captions validation set ($\approx 95.5\%$). We show additional statistics in the supplementary material.

3.3. Definition of the BISON Task

In the BISON task, the model is given two images and a text query that describes one of the two images and asked to select the correct image; see Figure 1. The model's performance is measured in terms of binary classification accuracy. We report the mean accuracy over the COCO-BISON data and refer to it as the BISON score. We only use COCO-BISON for evaluation, *i.e.*, we do not train systems on the annotations in the COCO-BISON dataset.

Existing text-based image retrieval and image captioning systems can be used to perform binary image selection. Doing so requires computing a "compatibility" score between the text query and the two images, and picking the image with the highest score. For image captioning systems, the compatibility score is generally defined as the log-likelihood of the text query given the image. Image retrieval systems naturally compute the compatibility score, *e.g.*, via an inner product of the image and text features.

4. BISON Evaluation of Text-Based Retrieval

We use the BISON task to evaluate captioning systems (in the supplementary material) and image retrieval methods (this section). We now evaluate four state-of-the-art image retrieval systems on the COCO-BISON dataset. The supplementary material describes our experimental setup in more detail.

4.1. Evaluated Retrieval Systems

We evaluate four systems for text-based image retrieval: (1) ConvNet+BoW, (2) ConvNet+Bi-GRU, (3) Obj+Bi-GRU, and (4) SCAN [10]. The *ConvNet+BoW* system represents the text query by averaging word embeddings over all words in the query, 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 the image-text compatibility score. The *ConvNet+Bi-GRU* system is identical to the previous system, but it follows [9] and uses a bi-directional GRU [5] to construct the text representation. The *Obj+Bi-GRU* system is similar to ConvNet+Bi-GRU but uses a Bi-

Dataset →	COCO-1K [8]				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 [9]	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 [10]	52.35	84.44	67.00	92.62	84.94
SCAN t2i [10]	**54.10**	**85.58**	**67.50**	**92.98**	**85.89**

Table 3: **Performance of text-based image retrieval systems:** Recall@k (with $k = 1$ and $k = 5$) of caption-based image retrieval and image-based caption retrieval on the COCO-1K dataset (left) compared to the BISON accuracy on the COCO-BISON dataset (right). See text for details.

GRU to aggregate image-region features (spatial ConvNet features or object proposal features) and construct the image representation. Finally, *SCAN* [10] 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 retrieval systems are trained to minimize a max-margin loss.

Following the current state-of-the-art in image retrieval [10], all systems use the top 36 object proposal features produced by a Faster R-CNN model [12] with a ResNet-101 backbone that was trained on the ImageNet and Visual Genome datasets.

4.2. Results

Table 3 presents the BISON accuracy of the text-based image 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 [8]. In line with prior work [10], we find that the SCANt2i system outperforms the competing systems in terms of all quality measures.

We observe that the ranking of retrieval systems in terms of BISON accuracy is similar to that in terms of retrieval measures. However, BISON provides a more reliable error measure because it does not erroneously consider correct retrievals to be incorrect just because another image happened to be labeled as the positive image for that query. This is reflected in the fact that the BISON score of all systems is higher than their recall@1, and implies that BISON scores are more reliable. We expect that the reliability of evaluation measures becomes more important as the quality of text-to-image matching systems increases.

5. Discussion

We proposed binary image selection (BISON) as an alternative task for evaluating the performance of systems that relate visual and linguistic content. We showed that BISON solves the issues of text-based image retrieval tasks that erroneously assume that all unlabeled images are negative examples for the text query. Compared to text-based image retrieval, BISON has the advantage that the evaluation is more reliable, easily interpretable, and focuses more on "fine-grained" visual content. Our evaluation of captioning models shows that BISON measures different characteristics of models compared to captioning measures like CIDEr. We hope that BISON fosters the development of systems that go beyond coarse-level matching of images and text.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 1, 2

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. 2018. 2

[3] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005. 2

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2, 3

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*, 2014. 3

[6] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013. 2

[7] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016. 2

[8] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 4

[9] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS Workshop*, 2014. 3, 4

[10] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. *ECCV*, 2018. 2, 3, 4

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 2

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4

[13] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. *CVPR*, 2019. 2

[14] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 1, 2

[15] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *T-ACL*, 2014. 3