Visual Semantic Relatedness Dataset for Image Captioning

Ahmed Sabir[1] Francesc Moreno-Noguer[2] Lluís Padró [1]
[1] Universitat Politècnica de Catalunya, TALP Research Center, Barcelona, Spain
[2] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain

Abstract

Modern image captioning system relies heavily on extracting knowledge from images to capture the concept of a static story. In this paper, we propose a textual visual context dataset for captioning, in which the publicly available dataset COCO Captions [30] has been extended with information about the scene (such as objects in the image). Since this information has a textual form, it can be used to leverage any NLP task, such as text similarity or semantic relation methods, into captioning systems, either as an end-to-end training strategy or a post-processing based approach.[1]

1. Introduction

Caption generation is a task that lies at the intersection of computer vision and natural language processing. This task aimed to generate a synthetic language description for a given image. Recently, Transformer [40] has become the new standard for image caption generation systems [10,20,27,45]. However, most diverse image captioning systems employ visual context information to generate accurate synthetic caption descriptions from an image. Early work, Fang *et al*. [14] use visual information from the image to build a caption re-ranking system via multimodal similarity. Another work, Wang *et al*. [42] focus on the importance of object information in images, such as frequency count, size, and position. Similarly, Coronia *et al*. [9] employ object information to control the caption generation as a visual grounding task. Gupta *et al*. [16] propose a contrastive learning based approach via language modeling and object information for weakly supervised phrase grounding in image captioning systems. Zhang *et al*. [44] explore semantic coherency in image captioning by aligning the visual context to the language graph, which results in capturing both the correct linguistic characteristics and visual relevance. Most recently, Sabir *et al*. [38] propose a belief revision based visual relatedness score that re-ranks the most visually related caption using the object information.

[1]Our dataset is available at https://github.com/ahmedssabir/Textual-Visual-Semantic-Dataset

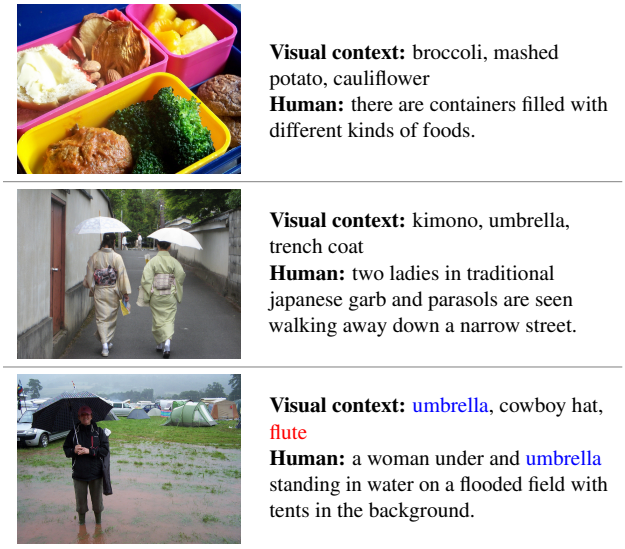

Figure 1. Examples of our proposed COCO based textual visual context dataset. (Top) the visual context associated with each image, (Bottom) the overlapping dataset in blue. We use out-of-the-box tools to extract visual concepts from the image. Figure 3 shows our proposed strategy to estimate the most closely related/not-related visual concepts to the caption description.

Learning the semantic relation between the text and its environmental context is an important task in computer vision *i.e*. a visual grounding task. While there are some publicly available visual context datasets for image captioning [1, 7, 30], none includes textual level information of the visual context in the image. Therefore, in this work, we propose a visual semantic relatedness dataset (Figure 1) for the caption pipeline, as our aim is to combine language and vision to learn textual semantic similarity and relatedness between the text and its related context from the image.

Our main contribution is this combined visual context dataset (visual contexts, caption), which provides the language and vision research community with the opportunity to use semantic similarity at the textual level between text and image to improve their results. In particular, we take a step further in pushing the limits of visual semantic con-

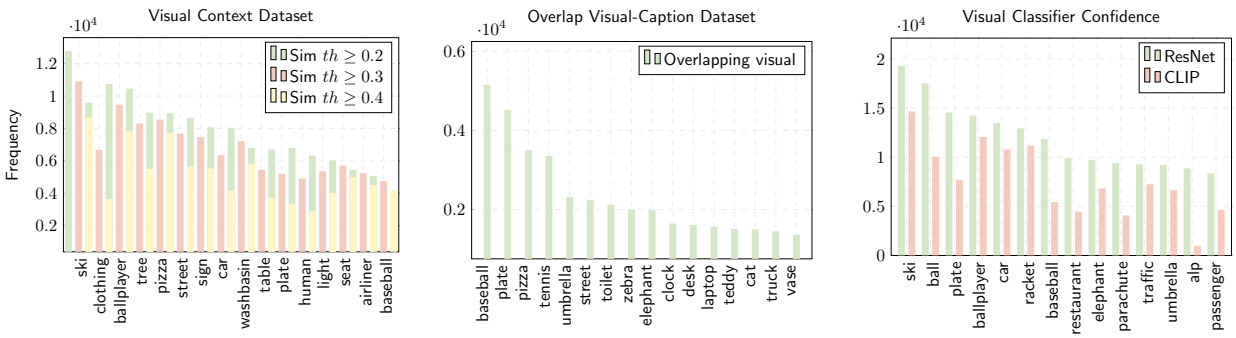

Figure 2. Visual semantic relatedness dataset. (**Left**) COCO-visual dataset: top frequency count of the extracted visual context from COCO Captions with semantic relatedness threshold with the human annotated caption. (**Middle**) COCO-overlapping dataset: top frequency count of the overlap visual context with human annotation. (**Right**) The figure shows the raw frequency count output of two visual classifiers (ResNet152 and CLIP). The result indicates that each classifier gave different degrees of confidence regarding the object in the image.

text data in image captioning by improving the data annotation pipeline used in [38] and introduce the visual semantic relatedness dataset. This dataset is based on the COCO dataset [30], and we further extend the dataset using out-of-the-box tools to extract the most closely related visual semantic context information from each image, as shown in Figure 2. Also, unlike the computer vision community, which tackles this problem by relying on visual features [11,28,32], our approach relies only on the textual information. Therefore, it can be used as an end-to-end or post-processing approach. In addition, we propose a similarity based re-ranking task, where the main concept is to learn a scoring function that assigns higher similarity to better candidate captions that correlate with its visual context from the same image.

2. Visual Context Information

To obtain the *visual context* from each image, we will use out-of-the-box classifiers to extract the image context information. We extract two kinds of context information: objects and scenarios present/seen in the image.

ResNet-152 [17]. A residual or shortcut connection-based deep network that relies heavily on batch normalization. The shortcut is known as a gated recurrent unit that is able to train a deeper network while maintaining less complexity.
CLIP [35]. (Contrastive Language-Image Pre-training) This is a pre-trained model with contrastive loss where the pair of image-text needs to be distinguished from randomly selected sample pairs. CLIP uses available resources across the Internet without human annotation of 400M pairs. CLIP achieves state-of-the-art performance in a wide range of image classification tasks in zero-shot learning.
Inception-Resnet FRCNN[2] **[19].** An improved variant of Faster R-CNN with the trade-off of better accuracy and fast inference via high-level features extracted from Inception-Resnet. Faster R-CNN has two stages: (1) a region proposal that suggests region-of-interest and (2) region-of-interest scoring. It is a pre-trained model that is trained on COCO categories with 80 object classes.

The objects extracted from all the pre-trained approaches mentioned above are obtained by extracting the top-3 object classes/categories from each classifier after filtering out no confidence instances via a probability threshold < 0.2.

[2]TensorFlow Object Detection API

3. Dataset

In this section, we first outline in more detail the existing datasets, and then we describe our proposed textual visual context information datasets.

3.1. Related Work

While there are a number of publicly available datasets for image captioning and visual context, none of them includes textual form (only in the form of a feature *e.g.* Visual genome [26] and Bottom-Up Top-Down feature [3]). In this section, we outline several publicly available datasets for image captioning tasks.
COCO [30]. This dataset (COCO Captions) contains more than 120K images, that are annotated by humans (five different captions per image). The most used data split by the language and vision community is provided by the Karpathy *et al.* [23], where 5K images are used for validation, 5K for testing, and the rest for training.
Novel Object Captioning [1]. A new dataset from the Open Images dataset [25] that extended for the image captioning task with the capability of describing novel objects which are not seen in the training set. The dataset consists of 15,100 images divided into validation and testing, 4,500 and 10,600, respectively. The images are grouped into subsets depending on their nearness to COCO classes.

Conceptual Captions 12M [7]. The most recent dataset and acquired image and text annotation from a web crawl. It contains around 12 million pairs automatically collected from the internet using relaxed filtering to increase the variety in caption styles.

3.2. Resulting Datasets

We rely on the COCO Captions dataset to extract the visual context information, as it is the most used by the language and vision community, and it was human annotated, as shown in Figure 1.

COCO-visual. It consists of 413,915 captions with associated visual context top-3 objects for training and 87,721 for validation. We rely on the confidence of the classifier to filter out non-existing objects in the image. For testing, we use VilBERT [32], with Beam search $k = 9$, to generate 45,000 captions with their visual context using the 5K Karpathy test split.

COCO-overlapping. Inspired by [42] that investigates the object count in image captioning. We also create an overlapping object with a caption as a dataset, as shown at the bottom of Figure 1. It consists of 71,540 overlap annotated captions and their visual context information.

Although we extract the top-3 objects from each image, we use three filter approaches to ensure the quality of the dataset (1) Threshold: to filter out predictions when the object classifier is not confident enough, and (2) semantic alignment with semantic similarity to remove duplicated objects. (3) a semantic relatedness score as a soft-label: to guarantee that the visual context and caption have a strong relation. In particular, we use Sentence RoBERTa [36] to give a soft label via cosine similarity[3] (*i.e.* the degree of visual relatedness) and then we use a threshold to annotate the final label (if $th \geq 0.2, 0.3, 0.4$ then [1,0]). Figure 2 shows the visual context dataset with different thresholds.

Figure 3 shows the proposed model to establish the visual context relatedness between the caption and the visual in the image. We omit higher threshold as the data becomes imbalanced with a more negative label.

Note that, all the textual visual contexts extracted by pre-trained models mentioned above have fast inference times, which makes them suitable for new task adoption. Therefore, we evade computationally hungry pre-computed features *e.g.* Bottom-Up Top-Down feature as it is too computationally expansive for our task.

4. Experiment

In this section, we describe the task and the experiment performed, and we compared the performance of our model against several existing baselines.

[3]Sentence-BERT uses a siamese network to derive meaningfully sentence embedding that can be compared via cosine similarity.

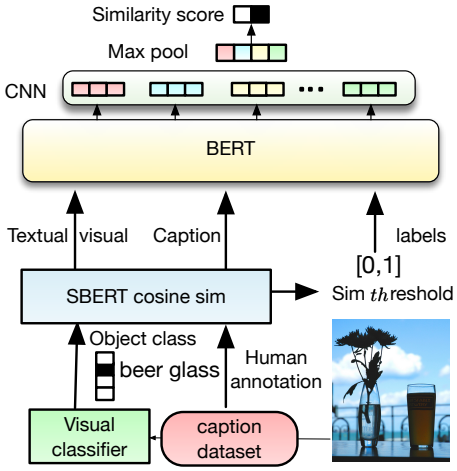

Figure 3. System overview. We propose an end-to-end system that (1) generates the visual semantic relatedness context dataset and (2) estimates the semantic relation between the candidate caption (provided by an off-the-shelf image captioning approach) and its environmental context in the image.

Task. To evaluate the dataset, we frame a re-ranking task, where the task is to re-rank the candidate captions produced by the baseline beam search using only similarity metrics. However, unlike previous works [14, 38], we rely only on similarity to re-rank the caption using semantic similarity[4] $sim(\text{visual}, \text{caption})$, and we, therefore, use top-3 multi-visual context information at the same time. We employ BERT/BERT-CNN based model as shown in Figure 3 to compute the similarity/relatedness score:

BERT [13]. BERT achieves remarkable results in semantic similarity, and we, therefore, fine-tune $\text{BERT}_{\text{base}}$ on the proposed dataset with a binary classifier, cross-entropy loss function [0,1] (related/not related).

BERT-CNN. To take advantage of the overlapping between the visual context and the caption, and to extract global information from each visual, we use the same BERT mentioned above as an embedding layer followed by a shallow CNN [24]. Let $X = \{w_1, \ldots, w_L\}$ be the sentence, where $w_i \in \mathbb{R}^D$ is the D dimensional BERT embedding of the i-th word vector in the sentence, while L denotes the sentence length. We pass the sentence X through a Kernel $\mathbf{f} \in \mathbb{R}^{K \times n \times D}$ that convolved over a window of n words with Kernel size K. By doing this operation, we generate local features of words n-gram fragments. The local feature of each i-th fragments is computed:

$$z_i = \mathrm{R}\left(\mathbf{f} * w_{i:i+n-1} + b\right) \quad (1)$$

[4]Semantic similarity is a more specific term than semantic relatedness. However, we here refer similarity to both semantic similarity and general semantic relatedness (*e.g. car* is similar to a *truck*, but is also related to *parking*).

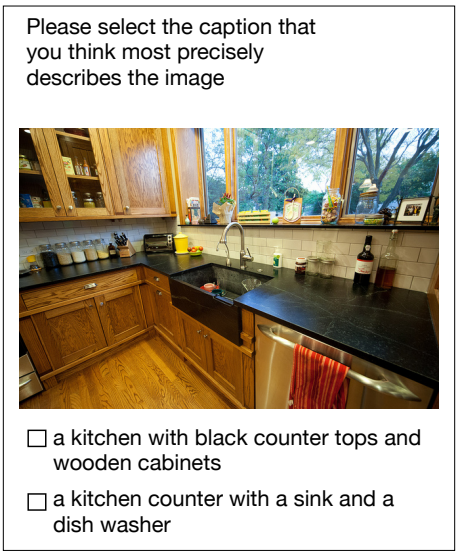

Figure 4. The user interface presented to our human subjects through the survey website asking them to re-rank the most descriptive caption candidates based on the visual information.

Model	B-4	M	R	C	S	B-S
VilBERT [32]	.330	.272	.554	1.104	.207	.9352
+ Best Beam	.351	.274	.557	1.115	.205	.9363
+$V_{\text{W-Object}}$ [14]	.348	.274	.559	1.123	.206	.9365
+V_{Object} [42]	.348	.274	.559	1.120	.206	.9364
+V_{Control} [9]	.345	.274	.557	1.116	.206	.9361
+SRoBERTa-sts (baseline)	.348	.272	.557	1.115	.204	.9362
+BERT $th = 0$	.345	.274	.558	1.117	.207	.9363
+BERT $th \geq 0.2$	.349	.275	.560	1.125	.207	.9364
+BERT $th \geq 0.3$	.351	.275	.560	1.127	.207	.9365
+BERT $th \geq 0.4$	.351	.276	**.561**	1.128	.207	**.9367**
+BERT-3LCNN $th = 0$	.350	.274	.558	1.121	.206	.9362
+BERT-3L-CNN $th \geq 0.2$	.349	.275	.559	1.128	.207	.9364
+BERT-3L-CNN $th \geq 0.3$	.350	.275	.560	**1.131**	.207	.9365
+BERT-3L-CNN $th \geq 0.4$	.350	.274	.559	1.124	.206	.9365
+BERT-CNN $th = 0$	.346	.275	.557	1.117	.207	.9361
+BERT-CNN $th \geq 0.2$	.349	**.277**	.560	1.128	**.208**	.9366
+BERT-CNN $th \geq 0.3$	**.352**	.275	.560	**1.131**	**.208**	.9366
+BERT-CNN $th \geq 0.4$	.348	.274	.560	1.123	.206	.9364

Table 1. Caption re-ranking performance results on the COCO Captions "Karpathy" test split. The result shows that the model benefits from having a threshold and n-gram extractor CNN over the baseline. The BERT-3L indicates that only the last BERT upper 3 layers are fine-tuned.

where * is the convolution operator, b is the bias, and $\mathrm{R}(\cdot)$ is the Rectified Linear Unit (ReLU) function [33]. By applying this convolution to all the sentence or text fragments, we obtain the corresponding feature map for the n-grams at all locations:

$$\mathbf{z} = [z_1, z_2, \ldots, z_{L-n+1}] \tag{2}$$

where the $\mathbf{z}$ or $\mathbf{z}^{(n)}$ is computed using BERT embedding, and each feature map has a size ($n = 3$). Then, all feature maps are finally concatenated with max pooling operator and then a sigmoid classification layer. We experimented, first, with fine-tuning the last BERT upper 3 layers (BERT-3L) to capture more semantic information with CNN. However, we gained more improvement in some metrics when fine-tuning the fully 12 layers in an end-to-end fashion. Note that, lower layers capture lexical information *i.e.* phrase-level [21], therefore, our approach also benefits from fine-tuning both lexical and semantic information.

Evaluation Metric. We use the official COCO offline evaluation suite, producing several widely used caption quality metrics: **B**LEU [34] **M**ETEOR [4], **R**OUGE [29], **C**IDEr [41], **S**PICE [2] and the semantic-similarity based metric **B**ERT**S**core (B-S) [43].

Human Evaluation. We conducted a small-scale human study to investigate human preferences over the visual re-ranked caption. We randomly selected 19 test images and gave the 12 reliable human subjects the option to chooses between two captions: (1) baseline and (2) similarity based visual ranker. We can observed that in around 50% of the cases, the human subject agreed with our re-ranking. Figure 4 shows the user interface presented to the human subjects, asking them to select the most diverse caption.

Baseline. We use visual semantic information to re-rank candidate captions produced by out-of-the-box state-of-the-art caption generators. We extract top-9 beam search candidate captions from general pre-trained vision-and-language model: VilBERT [32], fine-tuned on a total of 12 different vision and language datasets such as caption image retrieval and visual question answering.

Implementation. We apply different similarity based re-rankers as shown in Table 1. The re-rankers are similarity via fine-tuning BERT between the visual context and the candidate caption. The model is fined-tuned on each datasets that are labeled with different thresholds as shown in the Table 1, with a batch size 16 for 1 epoch with a learning rate 2e-5 and *max length* 50, we kept the rest of hyperparmeter settings as in the original implementation. For the BERT-CNN, the model is fined-tuned as end-to-end for five epochs.

5. Result and Discussion

We compared the performance of our model against several existing baselines that improve captions with object information. All baselines are trained on the same dataset (without any filtering *i.e.* threshold = 0), object based word re-ranking [14], an LSTM with object counter [42] and a language grounding based caption re-ranker [9].

The experiment consists of re-ranking the captions produced by the baseline pre-trained vision-and-language model VilBERT using only the similarity. In this experiment, each candidate caption is compared to multiple objects and concepts appearing in the image, and re-ranked according to

Model	Uniq	V	mB ↓	D1	D2	SB
VilBERT						
+ Best Beam	8.05	894	.899	.38	.44	.755
+$V_{\text{W-Object}}$ [14]	8.02	921	.899	.38	.44	.760
+V_{Object} [42]	8.03	911	.899	.38	.44	.757
+V_{Control} [9]	8.07	**935**	.899	.38	.44	.756
+BERT $th \geq 0.4$	7.98	794	.898	.38	.44	.759
+BERT-3L-CNN $th \geq 0.3$	8.06	903	.899	.38	.44	**.761**
+BERT-CNN $th \geq 0.2$	**8.15**	926	**.896**	.38	.44	.760
Human	9.14	3425	.750	.45	.62	NA

Table 2. **Diversity statistic**. *Div*-1 (D1) and *Div*-2 (D2) represent the ratio of unique unigram/bigram to the number of words in the caption. SBERT-sts (SB) indicates the average sentence level score between the caption and five human references. Also, we report the vocabulary size (V) and Uniq words per caption before and after re-ranking. Note that unlike the other metrics, lower *mBLEU* (mB) indicates more diverse re-ranked captions.

Model	B-4	M	R	C	S	B-S
VilBERT [32]	.330	.272	.554	1.104	**.207**	.9352
+ Best Beam	**.351**	**.274**	.557	1.115	.205	.9363
+$V_{\text{W-Object}}$ [14]	.348	.274	**.559**	**1.123**	.206	**.9365**
+V_{Object} [42]	.348	.274	**.559**	1.120	.206	.9364
+V_{Control} [9]	.345	.274	.557	1.116	.206	.9361
+S-BERT [36]	.348	.274	**.559**	**1.123**	.206	**.9365**
+S-BERT (distil)	.345	.273	.556	1.116	.206	.9360
+SimSCE [15]	.346	.273	.557	1.116	.206	.9362
+SimSCE (unsupervised)	.346	.274	.558	1.120	.206	.9364

Table 3. Performance results on the "Karpathy" test split via pre-tained model. All the pre-trained BERT models are $\text{RoBERTa}_{\text{Larage}}$ based models.

the obtained similarity scores. The results of our model and comparison against different baselines are reported in Table 1. The improvement is across all metrics with BERT except BLEU and SPICE. Therefore, we added CNN on the top of BERT to capture word-level global information and thereby we gained an improvement over word-level as shown in Figure 5.

Diversity Evaluation. We follow the standard diversity evaluation [12, 39]: (1) *Div*-1 (D1) the ratio of unique unigram to the number of word in the caption (2) *Div*-2 (D2) the ratio of unique bi-gram to the number of word in the caption, (3) *mBLEU* (mB) is the BLEU score between the candidate caption against all human captions (lower value indicates diversity) and finally (4) *Unique* words in the caption before and after re-ranking. Although, as shown in Table 2, the two first *Div* metrics are not able to capture the small changes, our results have lower mB and more *Unique* words per caption. Also, we use SBERT-sts[5] (SB) to measure the semantic diversity at the sentence level between the desired caption against the five human annotations. Figure 6 shows that SB (candidate caption against five human references average score) correlates more with humans than BERTscore.

[5]The model is out-of-the-box SBERT fine-tuned on the semantic textual similarity (sts) dataset [6].

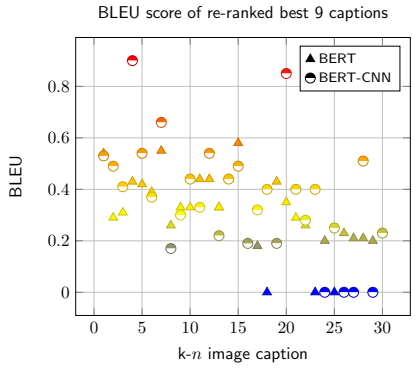

Figure 5. Result improvement of BERT on the BLEU score after adding CNN layer. Example with 30 images randomly selected from the "Karpathy" test split.

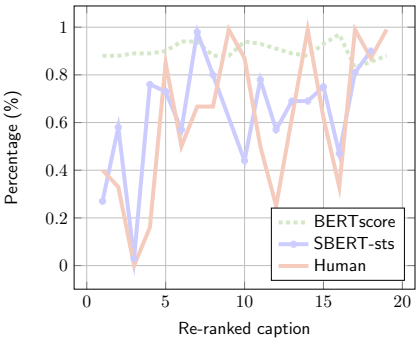

Figure 6. Comparison results between the small-scale human experiment and the automatic evaluation metrics BERTscore and sentence level SBERT-sts. The SBERT-sts as diversity metric correlates more with the five humans (average score) than BERT-score.

Experiments with Pre-trained Model. Although this approach is proposed to take the advantage of the dataset, we also investigate the use of an out-of-the-box based similarity re-ranker on the generated test set with visual context. For this, we use similarity to probability `SimProb` [5], but we only rely on similarity and the confidence of the classifier as:

$$\text{P}(w \mid c) = \text{sim}(w, c)^{\text{P}(c)} \tag{3}$$

where $\text{sim}(w, c)$ is the similarity between the visual contexts c and the caption w, and $\text{P}(c)$ is the visual classifier top-3 averaged confidence. We rely on two variations of RoBERTa [31] based model to compute the similarity: (1) SBERT that tuned on the STS-B dataset [6] and (2) a contrastive learning based semantic embedding SimSCE (supervise with NLI dataset [8] and unsupervised version). In particular, for the unsupervised approach, the model passes the sentence twice with dropout to obtain two embedding as positive pairs then the model predicts the positive one among other sentences

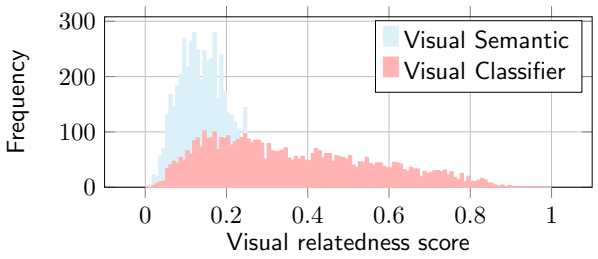

Figure 7. Visual relatedness score of our visual semantic model and the visual classifier based pre-trained model similarity. The fine-tuned similarity model relies more on the *semantic visual context* and thus surpasses out-of-the-box unstable relatedness (low/high) cosine similarity based *visual confidence* score.

Model	B-1	B-2	B-3	B-4	S	B-S
$\text{BLIP}_{\text{ViT-L}}$ [27]						
+ Best Beam (b=3)	.797	.649	**.514**	**.403**	**.243**	**.9484**
+BERT-CNN $th = 0$	.798	.646	.506	.392	.238	.9473
+BERT-CNN $th \geq 0.2$	.798	.647	.507	.393	.238	.9473
+BERT-CNN $th \geq 0.3$	.802	.651	.511	.397	.238	.9479
+BERT-CNN $th \geq 0.4$	**.804**	**.653**	.512	.398	.236	.9477

Table 4. Caption re-ranking results on the "Karpathy" test split using our visual semantic re-ranker on a state-of-the-art model (BLIP). Our model only improves the word level (B-1 and B-2) of each caption.

in the same mini-batch as negatives. The results are shown in Table 3, SBERT and SimSCE out-of-the-box models have the best results against the baselines in different metrics, especially in CIDEr, and slightly worse on BLEU-4 and SPICE.

Figure 7 shows a comparison results between the pre-trained model and our proposed similarity or visual semantic score. The pre-trained model relies on the visual classifier confidence and the unstable low/high relatedness score via pre-trained cosine similarity. Therefore, the pre-trained model struggles to associate the closest caption to its related visual context.

Experiments with State-of-the-Art Model. Although the main idea of the proposed task is to improve pre-trained models that were trained on low-mid-sized data (*i.e.* VilBERT is trained on 3.5M images) without any additional training. In this section, we experimented with the most recent state-of-the-art large pre-trained model Bootstrapping Language-Image Pre-training (BLIP) [27], a model of 125M pre-trained image (35.7x larger). We apply our visual semantic re-ranker on the best beam (B=3) suggested by the authors. As shown in Table 4 our best model improves only the word level BLEU-1/2 scores. In addition, the model improves the diversity of the selected caption as shown in Table 5, a lower *mBLEU* (mB), more vocabulary (V), and a higher human correlated score via SBERT-sts (SB).

Model	Uniq	V	mB ↓	D1	D2	SB
$\text{BLIP}_{\text{ViT-L}}$						
+ Best Beam (b=3)	**8.60**	1406	.461	.68	.80	.805
+BERT-CNN $th = 0$	8.49	**1532**	.458	.68	.80	.804
+BERT-CNN $th \geq 0.2$	8.48	1486	.458	.68	.80	.805
+BERT-CNN $th \geq 0.3$	8.41	1448	.458	.68	.80	**.806**
+BERT-CNN $th \geq 0.4$	8.30	1448	**.455**	.68	.80	.805
Human	9.14	3425	.375	.74	.84	NA

Table 5. **Diversity statistic**. Comparison results against the state-of-the-art pre-trained model. The result shows that our model has improved the baseline by re-ranking a diverse caption (lower *mBleu* (mB), more vocabulary (V), and a semantically correlated caption with the human references SBERT-sts (SB).

	Obj Gender Freq			ratio		
Visual	+ person	+ man	+ woman	m	w	to-m
clothing	3950	3360	1490	.85	.37	.69
footwear	2810	1720	220	.61	.07	.88
racket	1360	440	150	.32	.11	.74
surfboard	820	80	10	.09	.01	.88
tennis	140	200	60	1.4	.42	.76
motorcycle	480	40	20	.08	.04	.66
car	360	120	30	.33	.08	.80
jeans	50	240	70	4.8	1.4	.77
glasses	50	90	60	1.8	1.2	.60

Table 6. Frequency count of object + gender in the training dataset. The dataset, in most cases, has more gender-neutral `person` than men or women. The **m**en/**w**omen ratio is computed against `person`, and the gender bias ratio is estimated against men (**to**wards men) in the dataset.

Model	B-4	M	R	C	S	B-S
VilBERT [32]	.330	.272	.554	1.104	.207	.9352
+ Best Beam	.351	.274	.557	1.115	.205	.9363
$+\text{V}_{\text{W-Object}}$ [14]	.348	.274	.559	1.123	.206	.9365
$+\text{V}_{\text{Object}}$ [42]	.348	.274	.559	1.120	.206	.9364
$+\text{V}_{\text{Control}}$ [9]	.345	.274	.557	1.116	.206	.9361
+BERT-CNN $th \geq 0.3$	**.352**	.275	**.560**	1.131	**.208**	**.9366**
$+\text{V}_{\text{GN}}$ [46]	.350	.275	.559	1.128	.207	.9365
$+\text{Visual}_{\text{G-N}} + \text{Caption}_{\text{G-N}}$	.350	**.276**	**.560**	**1.132**	**.208**	**.9366**

Table 7. Performance results of our model against the best model in Table 1 with/without gender bias (**G**ender **N**eutral) on the "Karpathy" test split. The color red indicates when the model is worse than/equal to the baseline.

Bias in Visual Context. Another task that can benefit from the proposed dataset is investigating the contribution of the visual context to gender bias in image captioning. COCO Captions is a gender bias dataset towards men [18, 46], and our visual context dataset suffers from the same bias. However, the neutral gender dominates in most cases, as shown in Table 6. We follow zhao *et al.* [46] in calculating the gender bias ratio towards men as:

$$\frac{\text{count(obj, m)}}{\text{count(obj, m)} + \text{count (obj, w)}} \tag{4}$$

where **m**an and **w**oman refer to the visual in the image,

Visual context: fountain, sax, oboe ✗
Human: black and white of two women sitting on a marble looking bench one of them looking at camera holding and eating a watermelon wedge with another woman from back in a chair.

Visual context: parachute, volleyball, pole ✗
Human: a woman wearing a multi-colored striped sweater holds her arms up triumphantly as a kite flies high in the sky.

Figure 8. Limitation of the dataset. The model struggles with complex backgrounds and out-of-context objects.

and the **count** is the co-occurrence with the **obj**ect as pairs in the dataset. The ratio to *person* is computed as:

$$\frac{\text{count(obj, m/w)}}{\text{count(obj, person)}} \tag{5}$$

To investigate this further and to show how the balance data affects the final accuracy negatively, we replace each specific gender with gender-neutral (person/people) (*e.g.* a ~~man~~ *person* on a skateboard in a park). Then, we train our best model again, as shown in Table 7. The result as we expected, a lower accuracy, as in some cases specifying the gender influences the similarity score. For example, *a woman is putting makeup on another woman in a chair* is more human-like natural language than *a person is putting makeup on another person in a chair*. However, when making both the visual and the caption gender-neutral, the model achieves a stronger result. By having a gender neutral *person*, the result is better than having the wrong gender *man* or *woman* as the model overcomes the cases when the gender is not obvious.

Limitation. The drawback of this dataset is that the visual classifier struggles with complex backgrounds (*i.e.* wrong visual, object hallucination [37], *etc.*), as shown in Figure 8. These problems can be tackled by either relying on human annotation or using a more computationally expansive visual classifier or semantic segmentation based model. Another limitation is the low/high cosine label score (*i.e.* low relatedness context score), which leads to wrong annotations of the relation between the visual and the caption. For example, a *paddle* and *a man riding a surfboard on a wave.* have a low cosine score. We tackled this problem by adding multiple concepts at the same time to have more context to the sentence (*i.e.* caption).

6. Application

Visual Context based Image Search. One of the intuitive applications of this approach is the **V**isual **C**ontext based Image **S**earch (VCS). The model takes the visual context as an input *query* and attempts to retrieve the most closely related image via caption matching (*i.e.* semantic sentence matching as query-to-document matching).

Following the same procedure in this work: (1) we extract the visual context from the image with the visual classifier, then (2) the textual visual context is used, as a keyword for the semantic search, to extract the most closely related caption to its visual context, and (3) a sentence matching algorithm with cosine distance and semantic similarity (*e.g.* SBERT) is employed to re-rank the top-k semantically related caption, in the test set, to retrieve the image. The most direct approach to performing a semantic search is to extract the embedding from the last hidden layer after fine-tuning the model (*i.e.* VCS_{BERT})[6] and then using a k Nearest Neighbor

Figure 9. Visual Context based Image Search via **visual context from an image**. Examples show how the visual context is used to retrieve the image via caption information. The k**NN** is the original retrieved caption from the fine-tuned model and the **top-**k is the top re-ranked match generated caption from the Karpathy test split. Note that, using a single concept *query* results in more accurate retrieval than multiple concepts as in the *beer* example.

Model	R@1	R@5	R@10	R@15
VCS-k_1	.89	**.88**	**.87**	**.84**
VCS-k_2	**.90**	**.88**	.85	.83
VCS-k_3	**.90**	.87	.85	.83

Table 8. Retrieval results with top-k 3 visual context on the Karpathy test split. Results depicted in terms of Recall@K (R@K).

[6] Since the index representation is taken from the last hidden layer, no additional training of the model is required.

search (kNN) to retrieve the caption given the visual context. We adopt an efficient similarity search *extract search* using GPU directly with FAISS [22][7]. The *extract search* is a brute-force search that extracts the nearest neighbor with inner product with normalized length that is equal to the cosine similarity.

Table 8 shows that, without extra training, the model achieves good retrieval results with the top-k 3 visual context on the caption Karpathy test split. Figure 9 shows some successful cases of context-based retrieval. Also, we found that using a single concept *query* results in more accurate retrieval images than multiple concepts, as shown in the same figure with *beer* ~~glass~~ example.

Limitations. The limitations of this approach are: First, it is very sensitive to out-of-vocabulary word-to-caption. For example, for a rare *query* such as *lama*, the model randomly output words without any relation (*puck*, *pole* and *stupa*). Secondly, it relies on the quality of the classifiers (*i.e*. object and caption) to retrieve the related image. For instance, in Figure 9, the false positive *guitar* instead of *knife* and the false positive *fountain* with a correct retrieved image from caption description.

Conclusions

In this work, we have proposed a COCO-based textual visual context dataset. This dataset can be used to leverage any text-based task, such as learning the semantic relation/similarity between a visual context and a candidate caption, either as post-processing or end-to-end training. Also, we proposed two tasks and an application that can take advantage of this dataset.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *ICCV*, 2019. 1, 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 4

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 2

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACLW*, 2005. 4

[5] Sergey V Blok, Douglas L Medin, and Daniel N Osherson. Induction as conditional probability judgment. *Memory & Cognition*, 2007. 5

[6] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017. 5

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1, 3

[8] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017. 5

[9] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, 2019. 1, 4, 5, 6

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 1

[11] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021. 2

[12] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *CVPR*, 2019. 5

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 3

[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, John C, et al. From captions to visual concepts and back. In *CVPR*, 2015. 1, 3, 4, 5, 6

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simg of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 5

[16] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, 2020. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[18] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 6

[19] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 2

[20] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *CVPR*, 2019. 1

[21] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL*, 2019. 4

[22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 8

[7] An open source library for fast nearest neighbor retrieval in high dimensional spaces.

[23] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[24] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014. 3

[25] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 2

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 6

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004. 4

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[32] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 2, 3, 4, 5, 6

[33] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 4

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 4

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[36] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3, 5

[37] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018. 7

[38] Ahmed Sabir, Francesc Moreno-Noguer, Pranava Madhyastha, and Lluís Padró. Belief revision based caption re-ranker with visual semantic information. *arXiv preprint arXiv:2209.08163*, 2022. 1, 2, 3

[39] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*, 2017. 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[41] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 4

[42] Josiah Wang, Pranava Madhyastha, and Lucia Specia. Object counts! bringing explicit detections back into image captioning. *arXiv preprint arXiv:1805.00314*, 2018. 1, 3, 4, 5, 6

[43] Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020. 4

[44] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *AAAI*, 2021. 1

[45] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, et al. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *CVPR*, 2021. 1

[46] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017. 6