Visual Semantic Relatedness Dataset for Image Captioning

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Ahmed Sabir[1], Francesc Moreno-Noguer[2] and Lluís Padró[1]

[1]TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

- Learning the **semantic relation** between the text and its environmental visual context is an important task in computer vision *i.e.* a visual grounding task.
- While there are some publicly available visual context datasets for captioning **COCO** [30], **Novel Object Captioning** [1], and **Conceptual Captions 12M** [7] none includes textual level information of the visual context in the image.

Contributions

We propose a **visual semantic relatedness dataset** for the caption pipeline, as we aim to combine language and vision to learn textual semantic similarity and relatedness between the text and its related context. Also, we introduce two tasks and an application that can take advantage of this dataset.

Visual context: broccoli, mashed potato, cauliflower
Human: there are containers filled with different kinds of foods.
Sim score: 0.2910 ✓

Visual context: kimono, umbrella, trench coat
Human: two ladies in traditional japanese garb and parasols.
Sim score: 0.1444 ✗

Visual context: sealyham terrier, toy, poodle
Human: a white dog has a purple frisbee in its mouth.
Sim score: 0.4511 ✓

Visual context: umbrella, cowboy hat, flute
Human: a woman under and umbrella standing in water on a flooded field.
Sim score: 0.1756 ✗

Visual Semantic Datasets

We rely on COCO-Captions dataset to extract the visual context. We employ visual classifiers to extract visual information from each image.

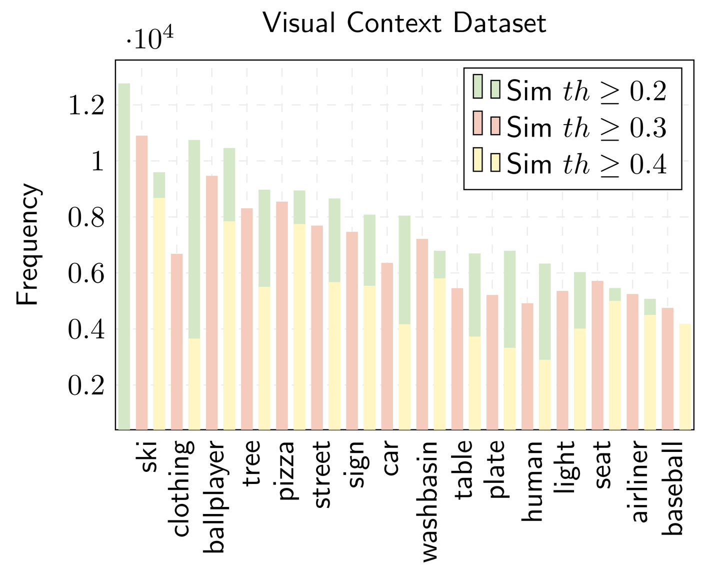
- **ResNet-152** [17]. To extract visual from 1000 ImageNet classes.
- **Inception-Resnet FRCNN** [19]. To extract object from COCO 80 categories.
- **CLIP** [35]. To extract out-of-domain classes.

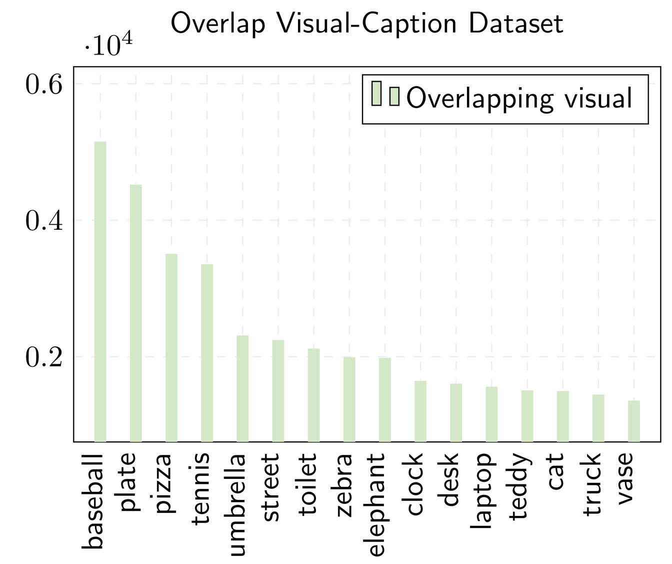
We extract the **top-3 objects** from each image, and we employ three **filter approaches** to ensure the quality of the dataset:

- **Threshold** to filter out predictions where the classifier is not confident enough.
- **Semantic Alignment** with semantic similarity to remove duplicated objects.
- **Semantic Relatedness** th**reshold via SentenceBERT-sts cosine similarity as a** *Sim* **soft-label** to guarantee that the visual context and caption have a strong relation. SBERT [36] is fine-tuned on semantic textual similarity task [6].

COCO-visual. It consists of 413,915 captions with associated visual context top-3 objects for training and 87,721 for validation.

COCO-overlapping. An overlapping object with a caption as a dataset. It consists of 71,540 overlapped annotated captions and their visual context information.

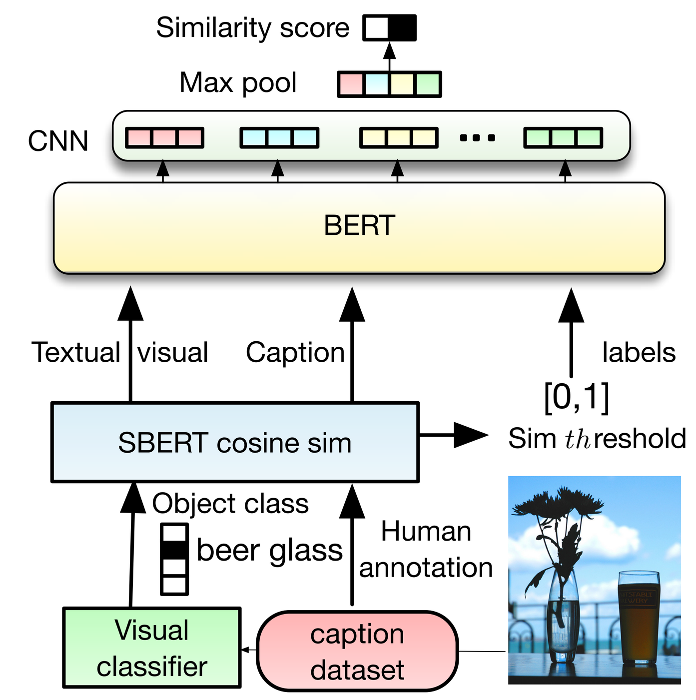


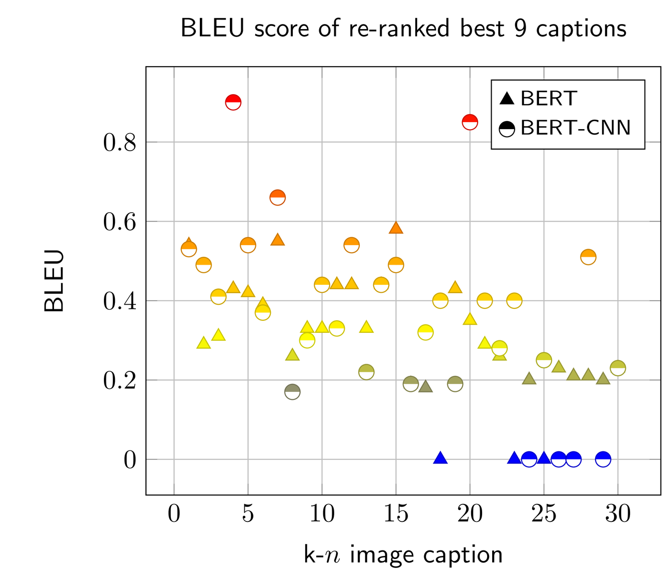


Proposed Method

We propose a strategy to estimate the most closely **related**/**not-related** visual concepts using the caption description.

BERT-CNN To take advantage of the overlapping between the visual context and the caption, and to extract global information from each visual, we fine-tuned BERT as an embedding layer and then we extract n-gram via shallow CNN [24]. Adding CNN improves learning the semantic correlation between the caption and its environmental visual context.

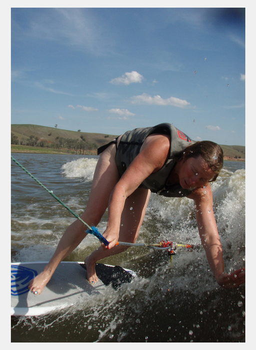




Task I: Caption Re-ranking

To evaluate the dataset, we frame a **re-ranking task**, where the task is to re-rank the caption hypotheses produced by the baseline beam search using **only similarity metrics**. We evaluate our model on two different pre-trained vision and language models in size (1) **VilBERT** [32] (trained on 3.5M images) and (2) **BLIP** [27] (trained on 124M images 35.7x larger).

Visual context: goblet, tree
VilBERT$_{\text{Beam}}$: a glass vase sitting on top of a table
VilBERT+Ours: a glass vase is sitting on a railing



Visual context: paddle, swimming trunks
BLIP$_{\text{Beam}}$: a woman riding a surfboard on top of a body of water
BLIP+Ours: a woman on a surfboard riding a wave

Examples of our proposed visual semantic re-ranker. The result shows that our model improves the baselines by selecting the most diverse caption using the visual context.

Model	B-4	M	R	C	S	BERTScore
VilBERT [32]	.351	.274	.557	1.115	.205	.9363
+V$_{\text{W-Object}}$ [14]	.348	.274	.559	1.123	.206	.9365
+V$_{\text{Object}}$ [42]	.348	.274	.559	1.120	.206	.9364
+V$_{\text{Control}}$ [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-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

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.

Task II: Gender Bias Evaluation

Another task that can benefit from the proposed dataset is investigating the contribution of the visual context to gender bias. Therefore, we also introduce a visual-to-caption Gender Neutral dataset.

	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

Frequency count of object + gender in the training dataset. The dataset, in most cases, has more gender-neutral *person* than gender bias toward men or women. The ratio is computed against *person* in the dataset. The dataset is similar to COCO, a gender bias dataset **to**ward men.

Application: Visual Context based Image Search

One of the intuitive applications of this approach is the **Visual Context based Image Search (VCS)**. The model takes the visual context as an input query and attempts to retrieve the most closely related image via caption matching.

Query	Visual	R@ Caption	R@10	R@ Image
	zebra	kNN: there is a adult zebra and a baby zebra in the wild top-k: a zebra and a baby in a field	100	
	pizza	kNN: a couple of people are eating a pizza top-k: a group of people sitting at a table eating pizza	90	
	✗ fountain	kNN: a fountain of water gushes in the middle of a street top-k: a fire hydrant spraying water onto the street	100	

Limitations

(1) The SBERT cosine soft-label is very sensitive to short/less diverse captions (due to the less sentence context), which leads to wrong annotations of the (visual, caption) relation, and (2) the visual classifiers struggle with complex backgrounds.

Visual context: fountain, sax, oboe ✗
Human: black and white of two women sitting on a marble looking bench one.

Visual context: parachute, volleyball, pole ✗
Human: a woman wearing a multi-colored striped sweater holds her arms.

Conclusion

- We have proposed a COCO-based textual visual semantic 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.
- Our dataset and code are publicly available on Github through this QR