Visual Semantic Relatedness Dataset for Image Captioning

Ahmed Sabir[‡], Francesc Moreno-Noguer[†], Lluís Padró[‡]

[‡] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain

[†] Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain

CVPR 2023 O-DRUM Workshop

Introduction

Motivation

- Learning the **semantic relation** between text/object and its surrounding environmental visual context is a crucial task in visual grounding (*i.e.* mapping the visual data to higher-level knowledge).
- Although there are datasets available for image captioning, such as COCO[30], Nocaps[1], and CC[7], none of them incorporate **textual-level information of the visual context present in the image.**

Human: there are containers filled with different kinds of foods.

Human: two ladies in traditional japanese garb and parasols are seen walking away down a narrow street.

Human: a white dog has a purple frisbee in its mouth.

Human: a woman under and umbrella standing in water on a flooded field with tents in the background.

Introduction

Contribution

- We propose a **visual semantic relatedness dataset** for the caption pipeline, as we aim to combine L&V in order to learn textual semantic similarity and relatedness between the visual and its related context.
- For each image, we extract three types of visual information: (1) 1K ImageNet classes ResNet[17], 2) COCO 80 categories Inception-ResNet FRCNN[19], and 3) CLIP[35] for rare/out-of-domain classes.

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 are seen walking away down a narrow street.
Sim score: 0.1444 ✗

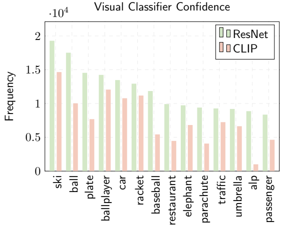
Visual context: sealyham terrier, toy, poodle
Human: a white dog has a purple frisbee in its mouth.
Sim score: 0.4511 ✓

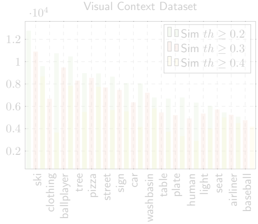
Visual context: umbrella, cowboy hat, flute
Human: a woman under and umbrella standing in water on a flooded field with tents in the background.
Sim score: 0.1756 ✗

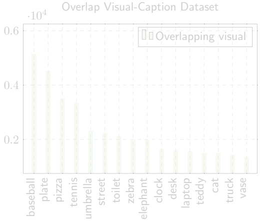
Introduction

Contribution

- ✓ To handle confidence variations among **visual classifiers** for objects in an image, we utilize the COCO-Captions dataset to extract the visual context. By leveraging the human caption as a reference, we establish a semantic relation to/with the objects in the image.
- ✓ We introduce two datasets: visual context and similarity soft-labels with the caption, and overlapping between objects and captions.



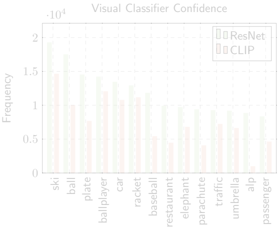


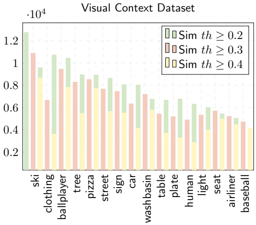


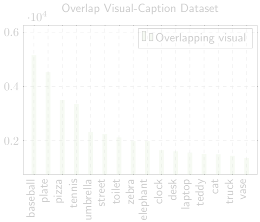
Introduction

Contribution

- ✓ To handle confidence variations among **visual classifiers** for objects in an image, we utilize the COCO-Captions dataset to extract the visual context. By leveraging the human caption as a reference, we establish a semantic relation to/with the objects in the image.
- ✓ We introduce two datasets: **visual context and similarity** soft-labels with the caption, and overlapping between objects and captions.



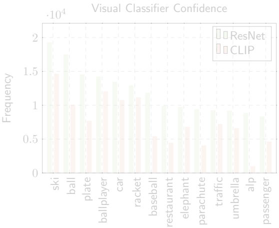


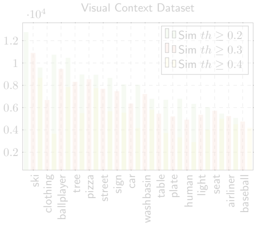


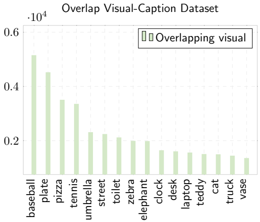
Introduction

Contribution

- ✓ To handle confidence variations among **visual classifiers** for objects in an image, we utilize the COCO-Captions dataset to extract the visual context. By leveraging the human caption as a reference, we establish a semantic relation to/with the objects in the image.
- ✓ We introduce two datasets: visual context and similarity soft-labels with the caption, and **overlapping** between **objects and captions**.







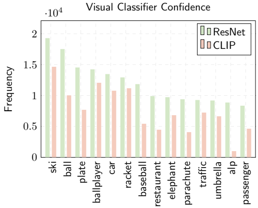
Dataset

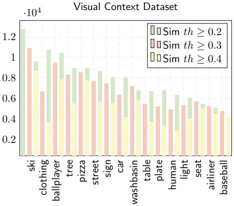
To ensure dataset quality, we apply three **filtering approaches** to the top-3 objects extracted from each image:

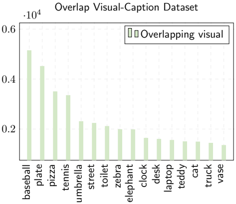
A **Threshold** to filter out predictions when the classifier is not confident enough.

B **Semantic Alignment** with word-level semantic similarity to remove duplicated objects.

C **Semantic Relatedness Score as *sim*ilarity soft-label** to guarantee that the visual context and caption have a semantic relation.



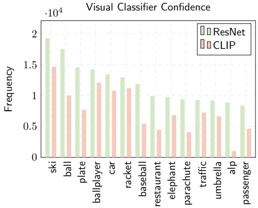


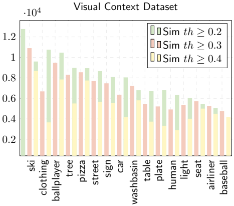


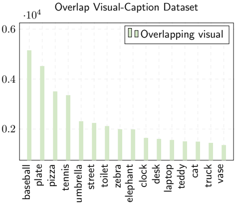
Dataset

To ensure dataset quality, we apply three **filtering approaches** to the top-3 objects extracted from each image:

- A **Threshold** to filter out predictions when the classifier is not confident enough.
- B **Semantic Alignment** with word-level semantic similarity to remove duplicated objects.
- C **Semantic Relatedness Score as *similarity soft-label*** to guarantee that the visual context and caption have a semantic relation.







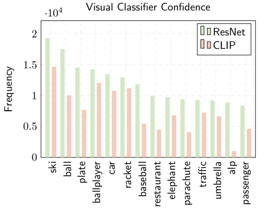
Dataset

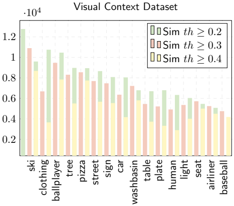
To ensure dataset quality, we apply three **filtering approaches** to the top-3 objects extracted from each image:

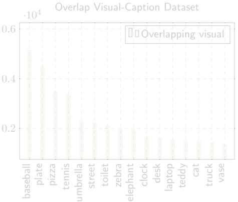
A **Threshold** to filter out predictions when the classifier is not confident enough.

B **Semantic Alignment** with word-level semantic similarity to remove duplicated/not related objects.

C **Semantic Relatedness Score as *sim*ilarity soft-label** to guarantee that the visual context and caption have a semantic relation.



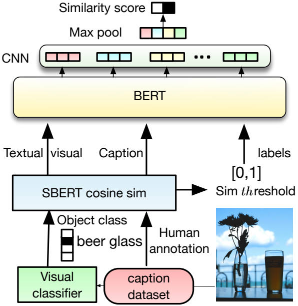




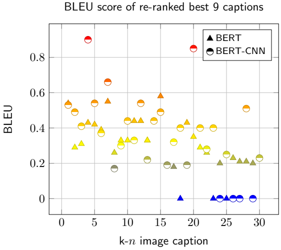
Proposed Method

We propose a strategy an end-to-end system to estimate the most closely related/not-related visual concepts based on the caption description (a).

BERT-CNN: to take advantage of the overlapping between the visual context and the caption, and to extract global information from the visual. (b) Improved BLEU score after adding CNN layer (Kim, 2014).



(a)



(b)

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.

- ViLBERT (trained on 3.5M images).
- BLIP (trained on 124M images 35.7x larger).

Model	B-4	M	R	C	S	BERTScore
ViLBERT [32]†	.351	.274	.557	1.115	.205	.9363
+$V_{\text{Multi-model Similarity}}$ [14]	.348	.274	.559	1.123	.206	.9365
+$V_{\text{Object Frequency}}$ [42]	.348	.274	.559	1.120	.206	.9364
+$V_{\text{Grounded Caption}}$ [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

† Jiasen Lu *et al.* 12-in-1: Multi-Task Vision and Language Representation Learning. CVPR2020.

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.

- ViLBERT (trained on 3.5M images).
- BLIP‡ (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

‡ Junnan Li *et al.* BLIP: Bootstrapping Lang/Image Pre-training for Unified VL Understanding and Generation. ICML2022

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.

The dataset primarily consists of a larger number of Gender-Neutral (person) instances rather than exhibiting gender bias towards men or women. As a result, we introduce a **Gender-Neutral** dataset.

	Object 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

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.

When using the object information, the relation of the visual-caption **G**ender-**N**eutral dataset will have a less negative impact on benchmark accuracy compared to gender balanced/reduced datasets.

Model	B-4	M	R	C	S	BERTScore
VilBERT [32]	.330	.272	.554	1.104	.207	.9352
+ Best Beam	.351	.274	.557	1.115	.205	.9363
$+\text{V}_{\text{Multi-model Similarity}}$ [14]	.348	.274	.559	1.123	.206	.9365
$+\text{V}_{\text{Object Frequency}}$ [42]	.348	.274	.559	1.120	.206	.9364
$+\text{V}_{\text{Grounded Caption}}$ [9]	.345	.274	.557	1.116	.206	.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
+ BERT-CNN $th \geq 0.3$						
$+ \text{V}_{\text{GN}}$ [46][‡] (reduced bias)	.350	.275	.559	1.128	.207	.9365
$+ \text{Visual}_{\text{GN}} + \text{Caption}_{\text{GN}}$	.350	**.276**	**.560**	**1.132**	**.208**	**.9366**

‡ Jieyu Zhao *et al*. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. EMNLP2017.

Application

One of the intuitive application of this approach is **V**isual **C**ontext based Image **S**earch (VCS), where the model utilizes the visual context as an input query. It then performs a similarity search to retrieve the most closely related image based on caption matching.

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

Application

One of the intuitive application of this approach is **V**isual **C**ontext based Image **S**earch (VCS), where the model utilizes the visual context as an input query. It then performs a similarity search[22] to retrieve the most closely related image based on caption matching[36].

Model	R@1	R@5	R@10	R@15
BERT-CNN $th \geq 0.3$				
+ VCS-k_1	.89	**.88**	**.87**	**.84**
+ VCS-k_2	**.90**	**.88**	.85	.83
+ VCS-k_3	**.90**	.87	.85	.83

[22] FAISS: Johnoson *et al.* Billion-scale similarity search with GPUs. arxiv: 1702.08734, 2017

[36] Reimers *et al.* Sentence Embeddings using Siamese BERT-Networks. EMNLP2019

Conclusion

Contributions

- 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.
- Also, we introduced two tasks and an application that can take advantage of this dataset.

Thank You