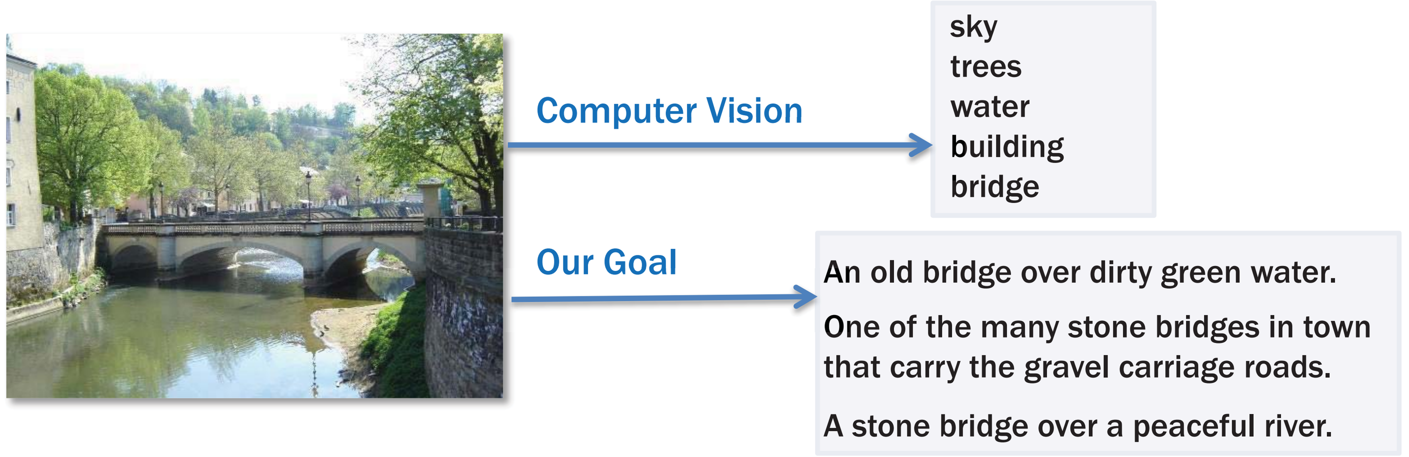
Im2Text: Describing Images Using 1 Million Captioned Photographs

Vicente Ordonez, Girish Kulkarni, Tamara L. Berg
Stony Brook University

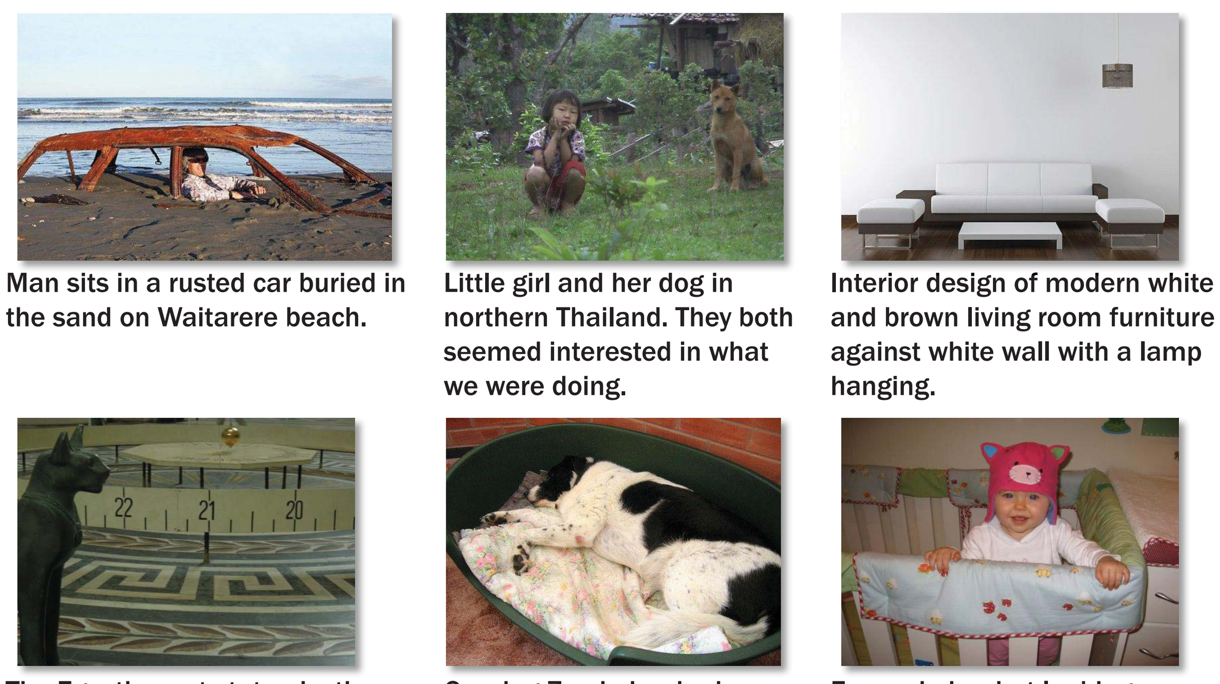
STONY BROOK
STATE UNIVERSITY OF NEW YORK

Contributions

- SBU Captioned Photo Dataset: A large novel data set containing *1 million* images from the web with associated captions written by people, filtered so that the descriptions are likely to refer to visual content. [http://tamaraberg.com/sbucaptions]
- A description generation method that utilizes global image representations to retrieve and transfer captions from our data set to a query image.
- A description generation method that utilizes both global representations and direct estimates of image content (objects, actions, stuff, attributes, and scenes) to produce relevant image descriptions.



SBU Captioned Photo Dataset



Man sits in a rusted car buried in the sand on Waitarere beach.

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing.

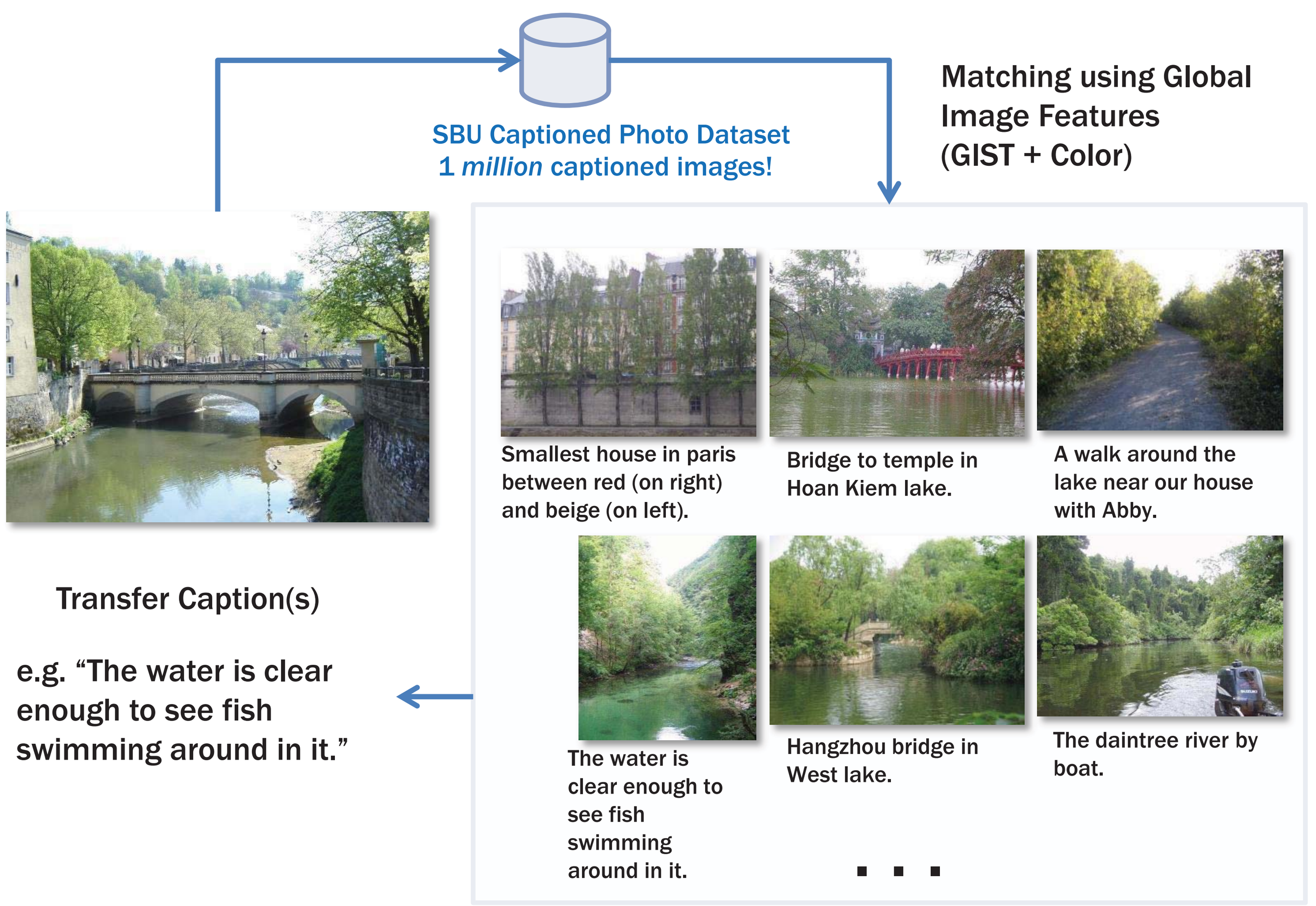
Interior design of modern white and brown living room furniture against white wall with a lamp hanging.

The Egyptian cat statue by the floor clock and perpetual motion machine in the pantheon.

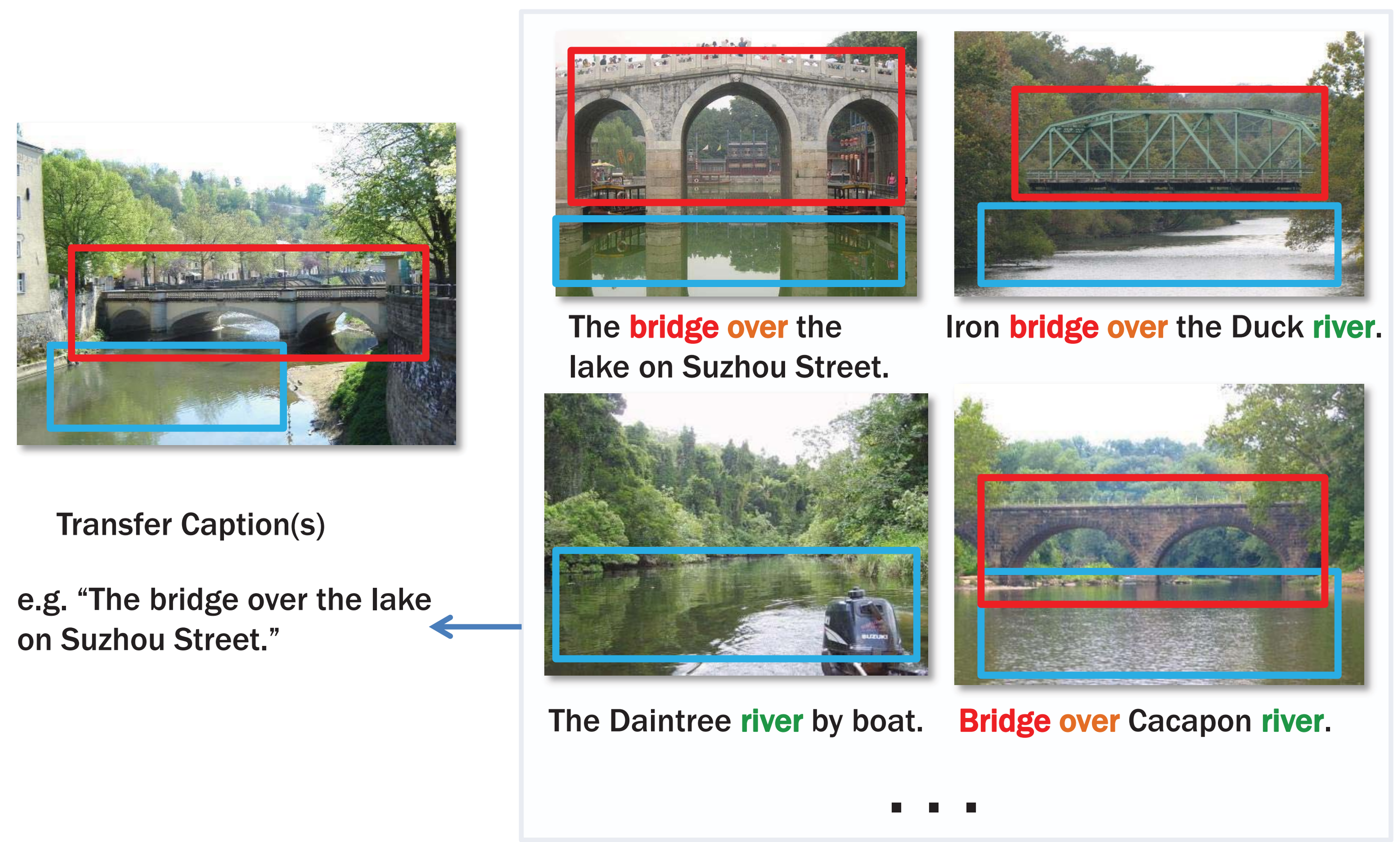
Our dog Zoe in her bed.

Emma in her hat looking super cute.

Method overview



Rerank retrieved images using high level content (captions, object detections, scene classification, stuff detections, people & actions)



High level information

Objects: 80 object categories using part-based deformable models and compute distances with objects detected in the query image based on visual attributes and raw visual descriptors.	**Stuff:** Detect stuff regions using a sliding window SVM scoring function with texton, color and geometric features as input. We determine similarity with the query image using product of SVM probabilities. (water, etc)
People/Actions: Detect people and pose using state-of-the-art methods and compute person similarity using an attribute based representation of pose.	**Scenes:** Train classifiers using global features for 26 common scene types and use the vector of classifier responses as a feature to compute similarity between images.

TFIDF: Rank the words in the returned set of image captions using their term-frequency inverse document frequency scores and follow a similar approach with the keywords for each object detection in the matching image set. As a result we obtain text-based TFIDF scores and object-detection-based TFIDF scores.

Dataset size

Past work on image retrieval has shown that small collections often produce spurious matches. Increasing data set size has a significant effect on the quality of retrieved global matches. Quantitative results also reflect this (see table at the bottom)

Good results

Amazing colours in the sky at sunset with the orange of the cloud and the blue of the sky behind.

A female mallard duck in the lake at Luukki Espoo

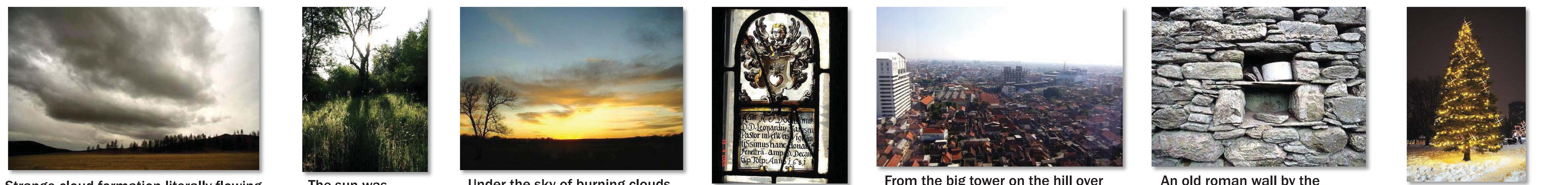
Fresh fruit and vegetables at the market in Port Louis Mauritius.

Tree with red leaves in the field in autumn.

One monkey on the tree in the Ourika Valley Morocco

Clock tower against the sky.

Reflection of the clear blue sky in the water.



Strange cloud formation literally flowing through the sky like a river in relation to the other clouds out there.

The sun was coming through the trees while I was sitting in my chair by the river

Under the sky of burning clouds.

Stained glass window in Eusebius church.

From the big tower on the hill over looking central Wakkanai.

An old roman wall by the tower of London.

A tree right around the corner from our house is this tree after the snow fell it was so beautiful.

Bad results

Incorrect objects

Kentucky cows in a field.

The cat in the window.

Incorrect context

Tree beside the river.

The sky is blue over the Gherkin.

Completely wrong

The boat ended up a kilometre from the water in the middle of the airstrip.

Water over the road.

BLEU score evaluation

Method	BLEU score
Global matching (1k)	0.0774 +- 0.0059
Global matching (10k)	0.0909 +- 0.0070
Global matching (100k)	0.0917 +- 0.0101
Global matching (1million)	0.1177 +- 0.0099
Global + Content matching (linear regression)	0.1215 +- 0.0071
Global + Content matching (linear SVM)	0.1259 +- 0.0060

Human evaluation

In addition, we propose a new evaluation task where a user is presented with two photographs and one caption. The user must assign the caption to the most relevant image. For evaluation we use a query image, a random image and a generated caption.

Caption used	Success rate
Original human caption	96.0%
Top caption	66.7%
Best from our top 4 captions	92.7%

The view from the 13th floor of an apartment building in Nakano awesome.

Please choose the image that better corresponds to the given caption: