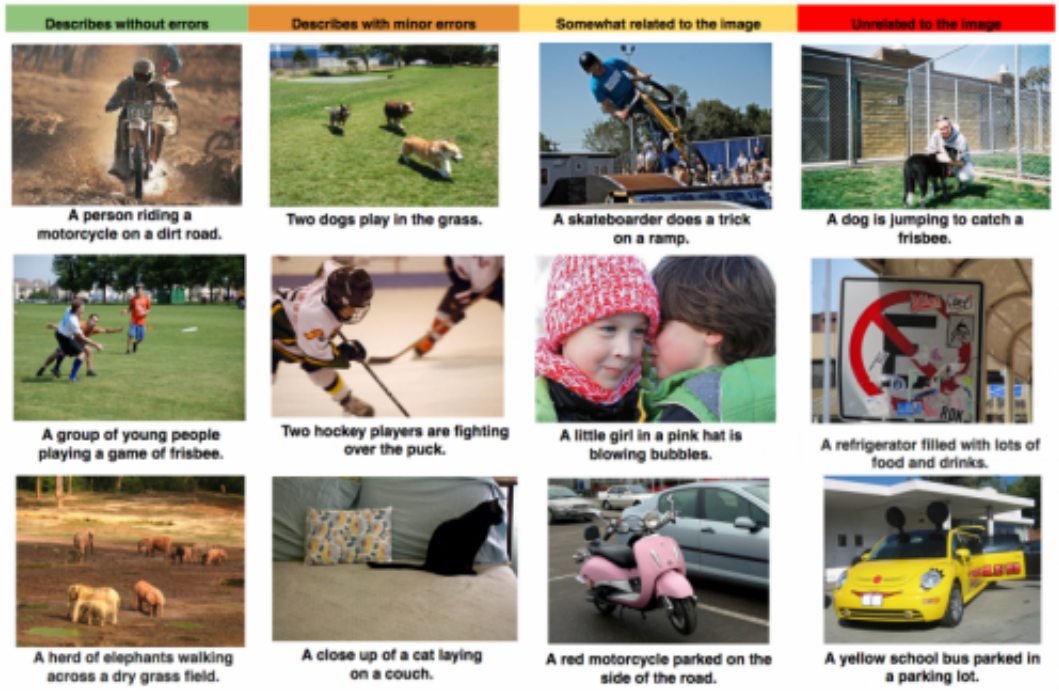
Image descriptions from computers show gains

November 18 2014, by Nancy Owano



A selection of evaluation results, grouped by human rating.

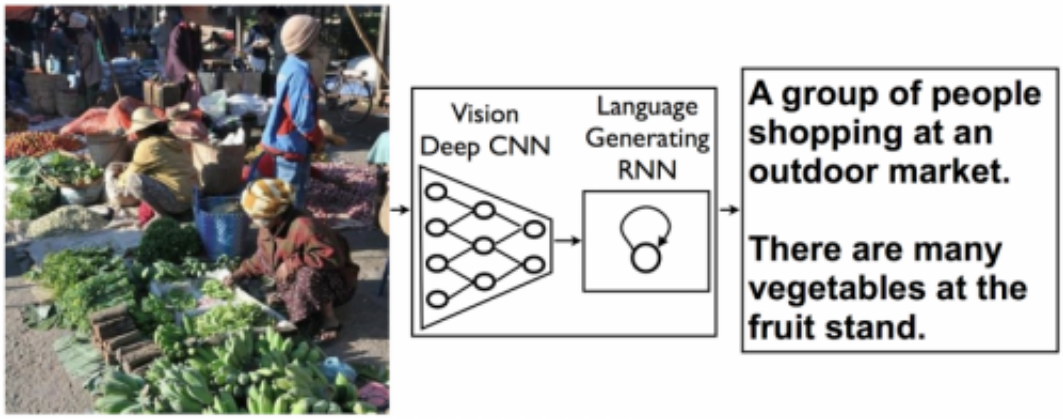
"Man in black shirt is playing guitar." "Man in blue wetsuit is surfing on wave." "Black and white dog jumps over bar." The picture captions were not written by humans but through software capable of accurately describing what is going on in images. At Stanford University, they have been working on Multimodal Recurrent Neural Architecture which generates sentence descriptions from images.

"I consider the pixel data in images and video to be the dark matter of the Internet," said Fei-Fei Li, associate professor and director of the Stanford Artificial Intelligence Laboratory, in The New York Times on Monday. She led research with Andrej Karpathy, a graduate student. "We are now starting to illuminate it." What is more, they are working on visual-semantic alignments between text and visual data. "Our alignment model learns to associate images and snippets of text." They provide examples of inferred alignments on their Stanford site. "For each image, the model retrieves the most compatible sentence and grounds its pieces in the image. We show the grounding as a line to the center of the corresponding bounding box. Each box has a single but arbitrary color."

Their technical report, "Deep Visual-Semantic Alignments for Generating Image Descriptions," takes us through their reasons for trying to improve on visual recognition. They wrote, "A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene. However, this remarkable ability has proven to be an elusive task for our visual recognition models. The majority of previous work in visual recognition has focused on labeling images with a fixed set of visual categories, and great progress has been achieved in these endeavors. However, while closed vocabularies of visual concepts constitute a convenient modeling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose."

They have taken on the challenge of trying to design a model rich enough to reason simultaneously about contents of images and their representation in the domain of natural language. "We first describe neural networks that map words and image regions into a common, multimodal embedding. Then we introduce our novel objective, which learns the embedding representations so that semantically similar concepts across the two modalities occupy nearby regions of the space."

Meanwhile, researchers at Google are making their own strides as well. In today's image-centric digital marketplace, architects cannot afford to live by images alone. "A picture may be worth a thousand words, but sometimes it's the words that are most useful—so it's important we figure out ways to translate from images to words automatically and accurately," blogged Google Research Scientists Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan on Monday.



The model combines a vision CNN with a language-generating RNN so it can take in an image and generate a fitting natural-language caption.

"People can summarize a complex scene in a few words without thinking twice. It's much more difficult for computers."

They said that they have developed a machine learning system that can automatically produce captions. "This kind of system could eventually help visually impaired people understand pictures, provide alternate text for images in parts of the world where mobile connections are slow, and make it easier for everyone to search on Google for images."

Writing in The New York Times on Monday, John Markoff noted

another advantage: "The advances may make it possible to better catalog and search for the billions of images and hours of video available online, which are often poorly described and archived."

More information: googleresearch.blogspot.com/20 … ousand-coherent.html
cs.stanford.edu/people/karpath … gesent/devisagen.pdf

© 2014 Tech Xplore

Citation: Image descriptions from computers show gains (2014, November 18) retrieved 3 August 2024 from https://techxplore.com/news/2014-11-image-descriptions-gains.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.