



Google Research Blog

The latest news from Research at Google

A picture is worth a thousand (coherent) words: building a natural description of images

Posted: Monday, November 17, 2014



1,950

Tweet

1,412

Me gusta

Posted by Google Research Scientists Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan

"Two pizzas sitting on top of a stove top oven"

"A group of people shopping at an outdoor market"

"Best seats in the house"

People can summarize a complex scene in a few words without thinking twice. It's much more difficult for computers. But we've just gotten a bit closer -- we've developed a machine-learning system that can automatically produce captions (like the three above) to accurately describe images the first time it sees them. 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.

Recent research has greatly improved [object detection](#), [classification](#), and [labeling](#). But accurately describing a complex scene requires a deeper representation of what's going on in the scene, capturing how the various objects relate to one another and translating it all into natural-sounding language.

Many efforts to construct computer-generated natural descriptions of images propose combining current state-of-the-art techniques in both [computer vision](#) and [natural language processing](#) to form a [complete image](#)

Research at Google

google.com/+ResearchatGoogle

vx, CS+vx



Seguir

+1

+ 873.406

[Labels](#)



Automatically captioned: "Two pizzas sitting on top of a stove top oven"

[Archive](#)[Feed](#)[Follow @googleresearch](#)

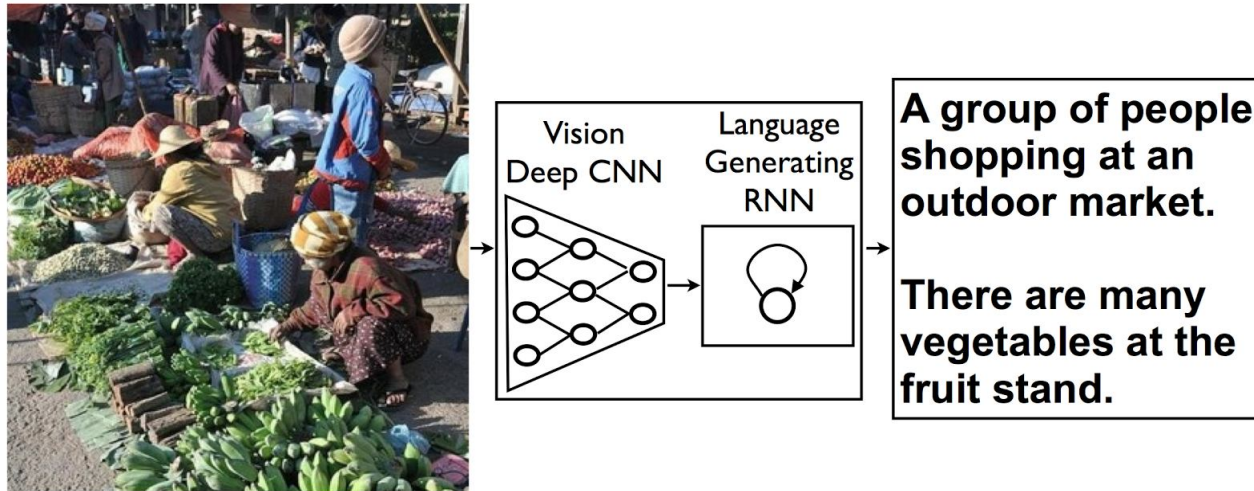
Give us feedback in our [Product Forums](#).

[description approach](#). But what if we instead merged recent computer vision and language models into a single jointly trained system, taking an image and directly producing a human readable sequence of words to describe it?

This idea comes from recent advances in [machine translation](#) between languages, where a [Recurrent Neural Network](#) (RNN) transforms, say, a French sentence into a [vector representation](#), and a second RNN uses that vector representation to generate a target sentence in German.

Now, what if we replaced that first RNN and its input words with a deep [Convolutional Neural Network](#) (CNN)

trained to classify objects in images? Normally, the CNN's last layer is used in a final [Softmax](#) among known classes of objects, assigning a probability that each object might be in the image. But if we remove that final layer, we can instead feed the CNN's rich encoding of the image into a RNN designed to produce phrases. We can then train the whole system directly on images and their captions, so it maximizes the likelihood that descriptions it produces best match the training descriptions for each image.












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.

Our experiments with this system on several openly published datasets, including Pascal, Flickr8k, Flickr30k and SBU, show how robust the qualitative results are -- the generated sentences are quite reasonable. It also performs well in quantitative evaluations with the [Bilingual Evaluation Understudy](#) (BLEU), a metric used in machine translation to evaluate the quality of generated sentences.

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. As the datasets suited to learning image descriptions grow and mature, so will the performance of end-to-end approaches like this. We look forward to continuing developments in systems that can read images and generate good natural-language descriptions. To get more details about the framework used to generate descriptions from images, as well as the model evaluation, read the full paper [here](#).

Labels: [Computer Vision](#), [Machine Learning](#), [Machine Translation](#), [Natural Language Processing](#)

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 <p>A person riding a motorcycle on a dirt road.</p>	 <p>Two dogs play in the grass.</p>	 <p>A skateboarder does a trick on a ramp.</p>	 <p>A dog is jumping to catch a frisbee.</p>
 <p>A group of young people playing a game of frisbee.</p>	 <p>Two hockey players are fighting over the puck.</p>	 <p>A little girl in a pink hat is blowing bubbles.</p>	 <p>A refrigerator filled with lots of food and drinks.</p>
 <p>A herd of elephants walking across a dry grass field.</p>	 <p>A close up of a cat laying on a couch.</p>	 <p>A red motorcycle parked on the side of the road.</p>	 <p>A yellow school bus parked in a parking lot.</p>

A selection of evaluation results, grouped by human rating.

463 comentarios



Añadir un comentario como Antonio Ferrandez

Mejores comentarios



Research at Google a través de Google+ · hace 2 días (editado) · Se ha compartido públicamente.

Building computer generated descriptions of images from raw pixels

"Two pizzas sitting on top of a stove top oven"

"A group of people shopping at an outdoor market"

Leer más (11 líneas) · Traducir

+437 1 · Responder

Ver las 16 respuestas



Awdaly Saleh · hace 17 horas

Absolutely

Traducir



Darvin Otero · hace 1 hora

When the API will be ready for developers?

Traducir



Chris Messina a través de Google+ · hace 2 días · Se ha compartido públicamente.

[Home](#)

[Older Post](#)



Company-wide

[Official Google Blog](#)

[Public Policy Blog](#)

[Student Blog](#)

Products

[Android Blog](#)

[Chrome Blog](#)

[Lat Long Blog](#)

Developers

[Developers Blog](#)

[Ads Developer Blog](#)

[Android Developers Blog](#)