Google Research Blog

The latest news from Research at Google

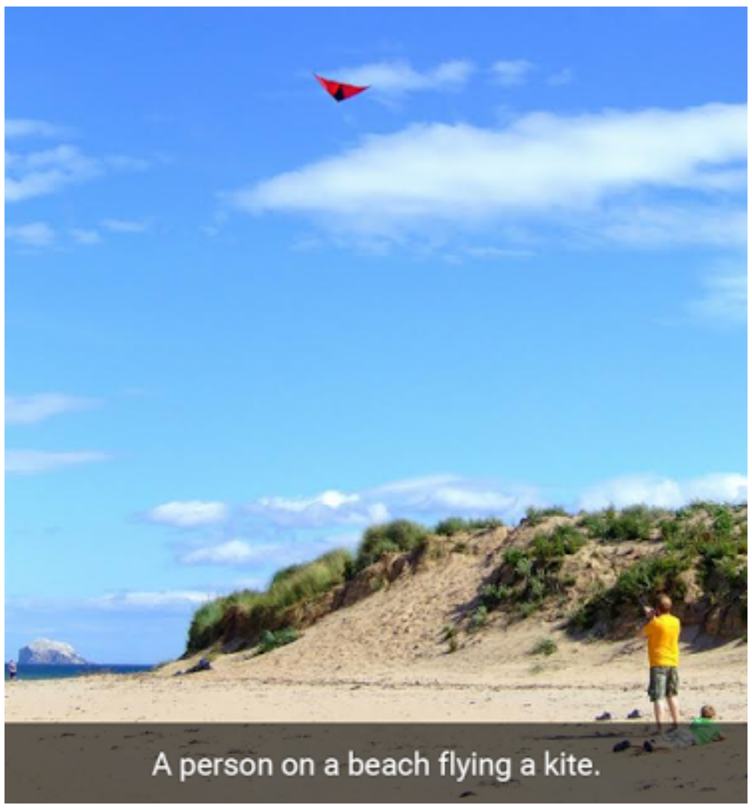
Show and Tell: image captioning open sourced in TensorFlow

Thursday, September 22, 2016

Posted by Chris Shallue, Software Engineer, Google Brain Team

In 2014, research scientists on the Google Brain team trained a machine learning system to automatically produce captions that accurately describe images. Further development of that system led to its success in the Microsoft COCO 2015 image captioning challenge, a competition to compare the best algorithms for computing accurate image captions, where it tied for first place.

Today, we're making the latest version of our image captioning system available as an open source model in TensorFlow. This release contains significant improvements to the computer vision component of the captioning system, is much faster to train, and produces more detailed and accurate descriptions compared to the original system. These improvements are outlined and analyzed in the paper *Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge*, published in IEEE Transactions on Pattern Analysis and Machine Intelligence.



Automatically captioned by our system.

So what's new?

Our 2014 system used the Inception V1 image classification model to initialize the image encoder, which produces the encodings that are useful for recognizing different objects in the images. This was the best image model available at the time, achieving 89.6% top-5 accuracy on the benchmark ImageNet 2012 image classification task. We replaced this in 2015 with the newer Inception V2 image classification model, which achieves 91.8% accuracy on the same task. The improved vision component gave our captioning system an accuracy boost of 2 points in the BLEU-4 metric (which is commonly used in machine translation to evaluate the quality of generated sentences) and was an important factor of its success in the captioning challenge.

Search blog ...

Labels

Archive

Feed

Google on

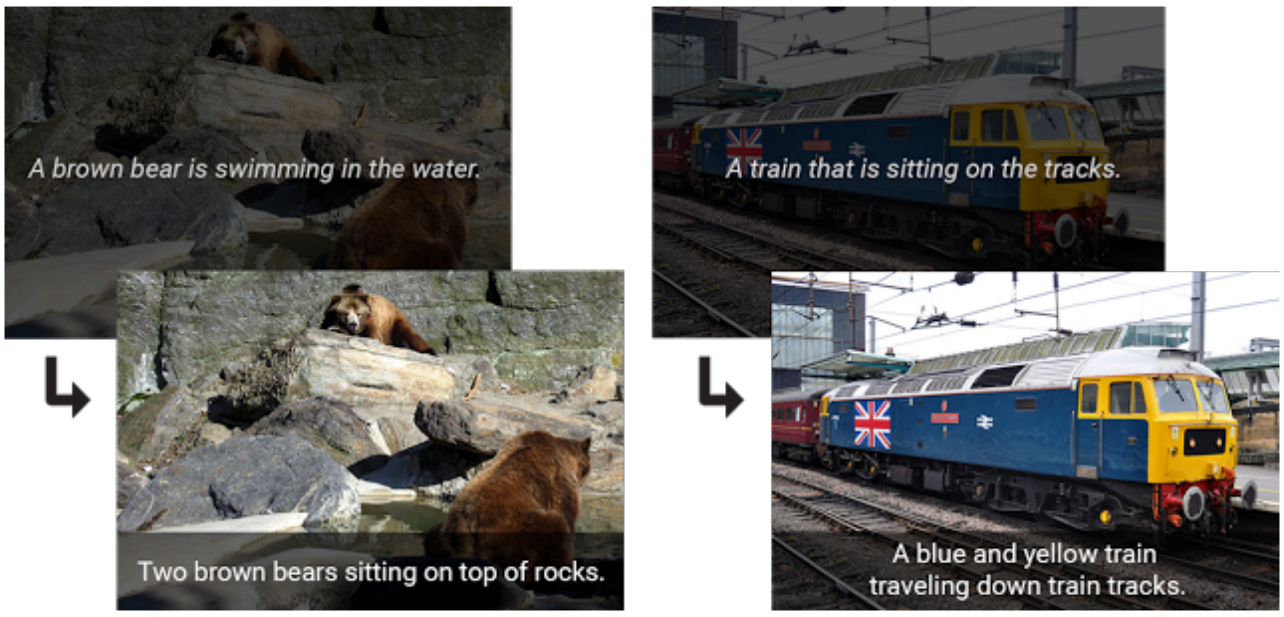
Follow @googleresearch

Give us feedback in our Product Forums.

Today's code release initializes the image encoder using the Inception V3 model, which achieves 93.9% accuracy on the ImageNet classification task. Initializing the image encoder with a better vision model gives the image captioning system a better ability to recognize different objects in the images, allowing it to generate more detailed and accurate descriptions. This gives an additional 2 points of improvement in the BLEU-4 metric over the system used in the captioning challenge.

Another key improvement to the vision component comes from *fine-tuning* the image model. This step addresses the problem that the image encoder is initialized by a model trained to *classify* objects in images, whereas the goal of the captioning system is to *describe* the objects in images using the encodings produced by the image model. For example, an image classification model will tell you that a dog, grass and a frisbee are in the image, but a natural description should also tell you the color of the grass and how the dog relates to the frisbee.

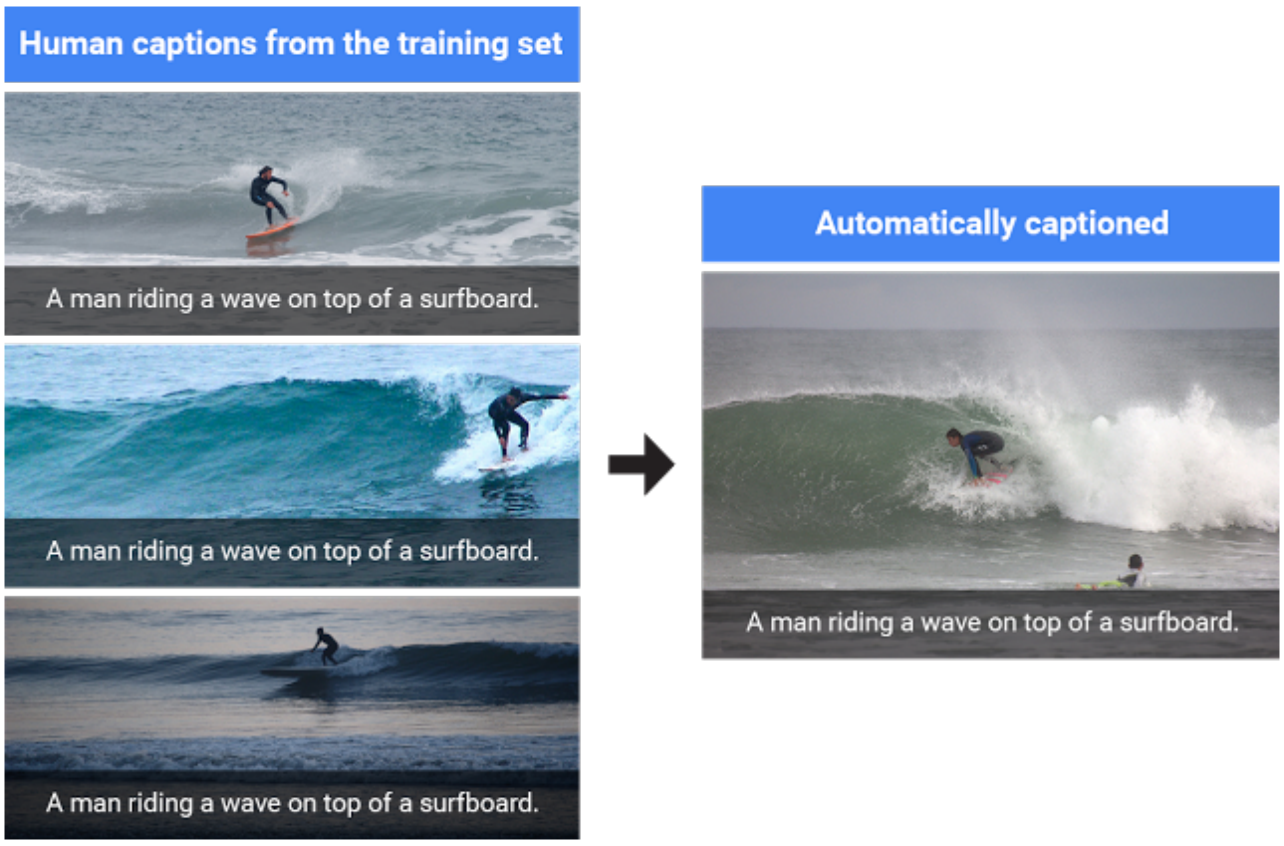
In the fine-tuning phase, the captioning system is improved by jointly training its vision and language components on human generated captions. This allows the captioning system to transfer information from the image that is specifically useful for generating descriptive captions, but which was not necessary for classifying objects. In particular, after fine-tuning it becomes better at correctly describing the colors of objects. Importantly, the fine-tuning phase must occur after the language component has already learned to generate captions - otherwise, the noisiness of the randomly initialized language component causes irreversible corruption to the vision component. For more details, read the full paper here.


A brown bear is swimming in the water.
Two brown bears sitting on top of rocks.
A train that is sitting on the tracks.
A blue and yellow train traveling down train tracks.

Left: the better image model allows the captioning model to generate more detailed and accurate descriptions. **Right:** after fine-tuning the image model, the image captioning system is more likely to describe the colors of objects correctly.

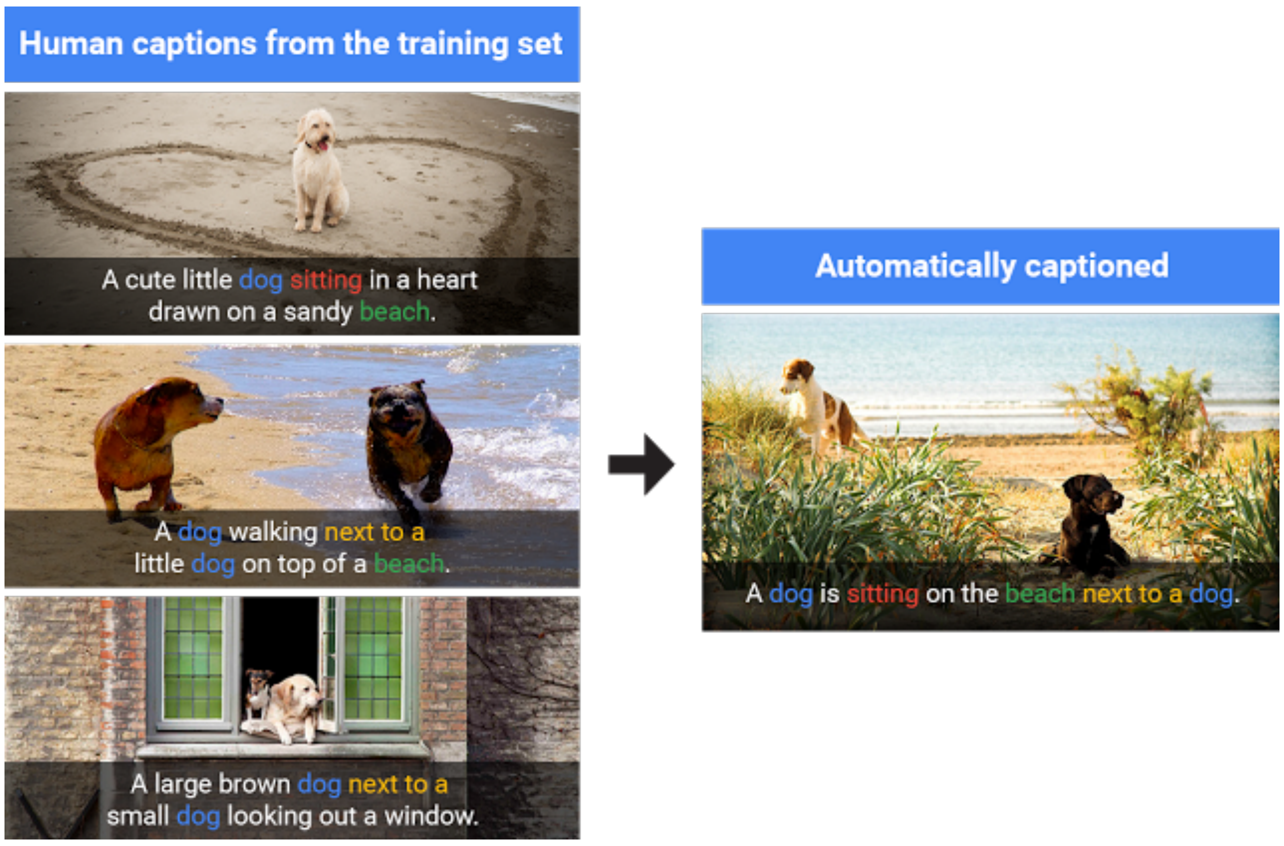
Until recently our image captioning system was implemented in the DistBelief software framework. The TensorFlow implementation released today achieves the same level of accuracy with significantly faster performance: time per training step is just 0.7 seconds in TensorFlow compared to 3 seconds in DistBelief on an Nvidia K20 GPU, meaning that total training time is just 25% of the time previously required.

A natural question is whether our captioning system can generate novel descriptions of previously unseen contexts and interactions. The system is trained by showing it hundreds of thousands of images that were captioned manually by humans, and it often re-uses human captions when presented with scenes similar to what it's seen before.


Human captions from the training set
A man riding a wave on top of a surfboard.
A man riding a wave on top of a surfboard.
A man riding a wave on top of a surfboard.
Automatically captioned
A man riding a wave on top of a surfboard.

When the model is presented with scenes similar to what it's seen before, it will often re-use human generated captions.

So does it really understand the objects and their interactions in each image? Or does it always regurgitate descriptions from the training data? Excitingly, our model *does indeed* develop the ability to generate accurate new captions when presented with completely new scenes, indicating a deeper understanding of the objects and context in the images. Moreover, it learns how to express that knowledge in natural-sounding English phrases despite receiving no additional language training other than reading the human captions.


Human captions from the training set
A cute little dog sitting in a heart drawn on a sandy beach.
A dog walking next to a little dog on top of a beach.
A large brown dog next to a small dog looking out a window.
Automatically captioned
A dog is sitting on the beach next to a dog.

Our model generates a completely new caption using concepts learned from similar scenes in the training set.

We hope that sharing this model in TensorFlow will help push forward image captioning research and applications, and will also allow interested people to learn and have fun. To get started training your own image captioning system, and for more details on the neural network architecture, navigate to the model's home-page here. While our system uses the Inception V3 image classification model, you could even try training our system with the recently released Inception-ResNet-v2 model to see if it can do even better!

148 条评论

以“Zhao Haijun”的身份发表评论

热门评论

Research at Google 通过 Google+ 7个月前 - 公开分享

Today, we're making the latest version of our image captioning system available as an open source model in **#TensorFlow** . This release contains significant improvements to the computer vision component of the captioning system, is much faster to train, and produces more detailed and accurate descriptions compared to the original system. Learn more in the link below!

· 翻译

+156 1 · 回复

查看所有 11 条回复

Parth Shah 6个月前

I am planning to modify model to use inception-resnet in this model but don't have enough GPU power to train. Can you suggest way to compensate it?

· 翻译

Caner Yarar 6个月前

Hi my name is Caner, I am specializing on AI & real artificial intelligence. This also covers Machine learning (everything which is related to machine learning). I am building an AIOS (Artificial Intelligence Operating System). Is the Google Brain team interested to support this kind of project ? You can find me on LinkedIn: Caner Yarar. Best wishes

· 翻译

Erik Jonker 通过 Google+ 7个月前（已编辑） - 公开分享

Image Captioning gets better

Good to see that image captioning gets even better. This part of the blog is the most interesting,

So does it really understand the objects and their interactions in each image? Or does it always regurgitate descriptions from the training data? Excitingly, our model does indeed

+44 1 · 回复

查看所有 5 条回复

Max Loh 7个月前 *+2*

+BennyOcean It was trained to maximize descriptiveness rather than elegance of prose. The overdone descriptiveness like "on top of" is what enabled it to make inferences about how to describe other objects in pictures without having been taught them explicitly.

· 翻译

Kwena Comment 7个月前

Simply got better with the new version

· 翻译

Artur Quaglio 7个月前 - 公开分享

It would be very neat if Google did something like the Image Labeler all over again... We'd have fun AND the AI would be trained :)

· 翻译

1 · 回复

Cristian Lorenzutti分享了此信息 通过 Google+ 7个月前 - 公开分享

Labels: Computer Vision , Google Brain , Machine Learning , TensorFlow

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

Google · Privacy · Terms