

# [NLP] 自己动手跑Google的Image Caption模型



ToeKnee (/u/5f50d2b82e05) [+ 关注](#)

2016.11.13 17:41\* 字数 1837 阅读 3660 评论 10 喜欢 5

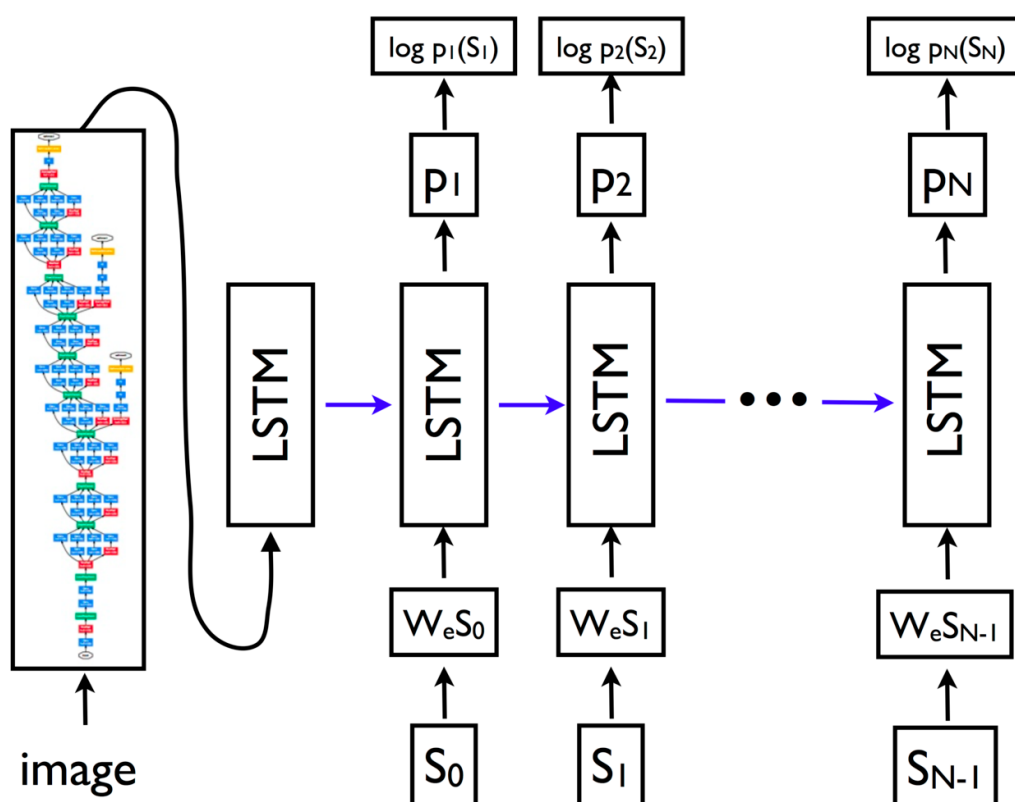
(/u/5f50d2b82e05)

两个月前Google公开了其之前在MSCOCO2015 Image Caption ([mscoco.org/dataset/#detections-challenge2015](http://mscoco.org/dataset/#detections-challenge2015))竞赛上夺得第一的Show&Tell (<http://arxiv.org/abs/1411.4555>)模型（与微软MSR基于DSSM的模型 (<http://arxiv.org/abs/1411.4952>)并列）基于TensorFlow的实现，最近在做这方面的工作，就试着跑了一下。代码工程在github上 (<https://github.com/tensorflow/models/tree/master/im2txt>)。RNN和LSTM的一些基本情况介绍可以参看这里：[\[NL系列\] RNN & LSTM 网络结构及应用](#) (<http://www.jianshu.com/p/f3bde26febed>)。

## Show&Tell/ im2txt

Google把公开之后的模型名称取为更像个工程名字的im2txt，其框架就像这张图：





图中  $s_{\{0\}}$  到  $s_{\{N\}}$  为生成的句子（包括开头和结尾各一个标识符）， $w_{\{e\}}s_{\{i\}}$  为第  $i$  个词对应的词向量，LSTM 的输出  $p_{\{i\}}$  是模型生成的句子中下一个单词（第  $i$  个）的概率分布。 $\log p_{\{i\}}(s_{\{i\}})$  代表位置  $i$  生成的单词正确性的log-likelihoods，这些值的总和的负数就是模型的最小化目标。

Google的模型采用了End-to-ends的思路，借用了机器翻译中的Encoder-Decoder (<https://arxiv.org/pdf/1406.1078v3.pdf>)框架（或者说是Google自己的Seq2Seq (<http://cs224d.stanford.edu/papers/seq2seq.pdf>)），通过一个模型直接将图像转换到句子。

机器翻译中Encoder-Decoder (Seq2Seq)模型的想法是，使用一个Encoder RNN读取源语言的句子，将其变换到一个固定长度的向量表示，然后使用 Decoder RNN将向量表示作为隐层初始值，产生目标语言的句子。

而im2txt的想法是，利用CNN在图片特征提取方面的强大能力，将Encoder RNN替换成CNN（im2txt中使用的是Google自己的Inception v3 (<http://arxiv.org/abs/1512.00567>)，模型在 ImageNet 分类任务上的准确率达到 93.9%，使得生成的图片描述的 BLEU-4 指标增加了 2 分），先利用CNN将图片转换到一个向量表示，再利用RNN将其转换到句子描述（采用beam search的方式，即迭代的在时刻 $t$ 时保存 $k$ 条最佳的句子片段用于生成 $t+1$ 时刻的词，生成 $t+1$ 时刻的词之后也只保存 $t+1$ 时刻的 $k$ 条最佳句子片段。代码中 $k$ 选择的是3，论文中说的是20，应该是照顾了人民群众的基础设施肯定不如Google的关系）。

在实现中，im2txt基于在ILSVRC-2012-CLS (<http://www.image-net.org/challenges/LSVRC/2012/>) 图片分类数据集上预训练好的CNN image recognition模型Inception v3 (<http://arxiv.org/abs/1512.00567>)，将其最后一个隐藏层作为Encoder RNN的输入，从而产生句



## Before Preparation

Whilst it is possible to run this code on a CPU, beware that this may be approximately 10 times slower.

## Preparation

1. Bazel (<https://bazel.build/versions/master/docs/install.html#install-with-installer>) : Bazel是Google开源的自动化构建工具，类似于Make的功能，用来编译构建tensorflow。链接中给出的是Bazel官方在Ubuntu14.04或15.04下的安装教程，如果使用Java7的话可以按照[这里](https://bazel.build/versions/master/docs/install.html#using-bazel-with-jdk-7-deprecated) (<https://bazel.build/versions/master/docs/install.html#using-bazel-with-jdk-7-deprecated>)的介绍稍作修改。
2. TensorFlow ([https://www.tensorflow.org/versions/master/get\\_started/os\\_setup.html](https://www.tensorflow.org/versions/master/get_started/os_setup.html)) : 注意安装的时候选择从源码编译的选项，按照支持GPU的步骤安装（所以首先要安装CUDA和CuDNN等）。中文版的教程 ([https://github.com/jikexueyuanwiki/tensorflow-zh/blob/master/SOURCE/get\\_started/os\\_setup.md%E4%BB%8E%E6%BA%90%E7%A0%81%E5%AE%89%E8%A3%85-](https://github.com/jikexueyuanwiki/tensorflow-zh/blob/master/SOURCE/get_started/os_setup.md%E4%BB%8E%E6%BA%90%E7%A0%81%E5%AE%89%E8%A3%85-))可能在命令的版本上有所区别，最新版的建议看英文官网 ([https://www.tensorflow.org/versions/master/get\\_started/os\\_setup.html](https://www.tensorflow.org/versions/master/get_started/os_setup.html))。
3. NumPy (<http://www.scipy.org/install.html>) : 基本上安装TensorFlow的时候都会装好。
4. Natural Language Toolkit (NLTK) : 用于NLP的开源python函数库。首先安装NLTK (<http://www.nltk.org/install.html>)，然后安装NLTK data (<http://www.nltk.org/data.html>)。



```
# Location to save the MSCOCO data.
MSCOCO_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/data/mscoco"

# Build the preprocessing script.
bazel build im2txt/download_and_preprocess_mscoco

# Run the preprocessing script.
bazel-bin/im2txt/download_and_preprocess_mscoco "${MSCOCO_DIR}"
```

等到输出下面这句话，数据集的准备就算完成一半了。

```
2016-09-01 16:47:47.296630: Finished processing all 20267 image-caption pairs in dat
```

剩下的一半是要把在ILSVRC-2012-CLS (<http://www.image-net.org/challenges/LSVRC/2012/>) 图片分类数据集上预训练好的*Inception v3* (<http://arxiv.org/abs/1512.00567>)模型下载下来。

This checkpoint file is provided by the TensorFlow-Slim image classification library (<https://github.com/tensorflow/models/tree/master/slim#tensorflow-slim-image-classification-library>) which provides a suite of pre-trained image classification models.

执行以下命令（注意可以到TensorFlow-Slim image classification library (<https://github.com/tensorflow/models/tree/master/slim#tensorflow-slim-image-classification-library>)看看最新的模型是什么，替换下面的 *inception\_v3\_2016\_08\_28.tar.gz*：

```
# Location to save the Inception v3 checkpoint.
INCEPTION_DIR="${HOME}/im2txt/data"
mkdir -p ${INCEPTION_DIR}

wget "http://download.tensorflow.org/models/inception_v3_2016_08_28.tar.gz"
tar -xvf "inception_v3_2016_08_28.tar.gz" -C ${INCEPTION_DIR}
rm "inception_v3_2016_08_28.tar.gz"
```

这个pre-trained模型只会在第一次执行训练时用到，im2txt每训练一段时间（默认的应该是迭代1024次）就会保存一次模型的checkpoint，之后的训练过程都会从checkpoint开始。

## Start Training

im2txt的模型训练分为两步，第一步的initial training会固定CNN部分（Inception V3）的参数，把其当作一个图像编码网络生成image embedding，参与训练的只有在Inception V3上增加的一层网络（用于将image embedding映射到LSTM的word embedding vector space），而LSTM部分的所有待训练参数在此都会参与训练。



```
# Directory containing preprocessed MSCOCO data.
MSCOCO_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/data/mscoco"

# Inception v3 checkpoint file.
INCEPTION_CHECKPOINT="${YOUR_ADDR_TO_IM2TXT}/im2txt/data/inception_v3.ckpt"

# Directory to save the model.
MODEL_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/model"

# Build the model.
bazel build -c opt im2txt/...

# Run the training script.
bazel-bin/im2txt/train \
  --input_file_pattern="${MSCOCO_DIR}/train-????-of-00256" \
  --inception_checkpoint_file="${INCEPTION_CHECKPOINT}" \
  --train_dir="${MODEL_DIR}/train" \
  --train_inception=false \
  --number_of_steps=1000000
```

在训练的同时可以执行evaluation，以在TensorFlow自带的TensorBoard上方便的查看当前训练情况。如果只有一个GPU的话没有办法同时在GPU上跑evaluation（内存不够），因此一般是在CPU上执行，可以在命令行中执行`export CUDA_VISIBLE_DEVICES=""`命令限制当前程序看不到CUDA设备。默认的evaluation每600秒执行一次，在从最初的Inception V3模型迭代5000次之后才会开始，这些参数和设置都可以通过查看evaluate.py的代码了解。

```
MSCOCO_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/data/mscoco"
MODEL_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/model"

# Ignore GPU devices (only necessary if your GPU is currently memory
# constrained, for example, by running the training script).
export CUDA_VISIBLE_DEVICES=""

# Run the evaluation script. This will run in a loop, periodically loading the
# latest model checkpoint file and computing evaluation metrics.
bazel-bin/im2txt/evaluate \
  --input_file_pattern="${MSCOCO_DIR}/val-????-of-00004" \
  --checkpoint_dir="${MODEL_DIR}/train" \
  --eval_dir="${MODEL_DIR}/eval"
```

然后就可以开启一个TensorBoard进程通过浏览器监控训练进度。

```
MODEL_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/model"

# Run a TensorBoard server.
tensorboard --logdir="${MODEL_DIR}"
```

## Generating Captions

其实在训练的过程中随时可以生成图片描述，只是效果并不好说（其实也不一定比迭代很久之后差！）



执行以下命令：

```
# Directory containing model checkpoints.
CHECKPOINT_DIR="${YOUR_ADDR_TO_IM2TXT}/im2txt/model/train"

# Vocabulary file generated by the preprocessing script.
VOCAB_FILE="${YOUR_ADDR_TO_IM2TXT}/im2txt/data/mscoco/word_counts.txt"

# JPEG image file to caption.
IMAGE_FILE="${YOUR_ADDR_TO_IM2TXT}/im2txt/data/mscoco/raw-data/val2014/${CHOICE_OF_I

# Build the inference binary.
bazel build -c opt im2txt/run_inference

# Ignore GPU devices (only necessary if your GPU is currently memory
# constrained, for example, by running the training script).
export CUDA_VISIBLE_DEVICES=""

# Run inference to generate captions.
bazel-bin/im2txt/run_inference \
  --checkpoint_path=${CHECKPOINT_DIR} \
  --vocab_file=${VOCAB_FILE} \
  --input_files=${IMAGE_FILE}
```

官方给出的sample如下：



COCO\_val2014\_000000224477.jpg: a man riding a wave on top of a surfboard .

Captions for image COCO\_val2014\_000000224477.jpg:

- 0) a man riding a wave on top of a surfboard . (p=0.040413)
- 1) a person riding a surf board on a wave (p=0.017452)
- 2) a man riding a wave on a surfboard in the ocean . (p=0.005743)



其实在我跑的时候大概迭代到200000次时同样是这张图生成的caption感觉比现在的第一条还要更合理一些，这就见仁见智了。

## If you want more

如果之前的训练你觉得已经足够久，或者生成的caption你觉得还需要进一步优化，或者你正在苦于怎么超过state-of-art，那就可以把CNN的参数也一起放进来训练了，执行以下命令：

```
# Restart the training script with --train_inception=true.
bazel-bin/im2txt/train \
  --input_file_pattern="${MSCOCO_DIR}/train-????-of-00256" \
  --train_dir="${MODEL_DIR}/train" \
  --train_inception=true \
  --number_of_steps=3000000 # Additional 2M steps (assuming 1M in initial training)
```

来自Google的温馨提醒：

Note that training will proceed much slower now, and the model will continue to improve by a small amount for a long time. We have found that it will improve slowly for an additional 2-2.5 million steps before it begins to overfit. This may take several weeks on a single GPU.

## A Little Thoughts

还能有什么感想呢，现在initial training都没跑完。

这里有个日文的report (<http://tensorflow.classcat.com/2016/11/10/tensorflow-image-caption/>)和本文内容差不多，可以参考，里面有从TensorBoard中截取出来的图像。

学术！笔记！（/nb/4332710）

举报文章 © 著作权归作者所有



ToeKnee (/u/5f50d2b82e05)

写了 22286 字，被 162 人关注，获得了 115 个喜欢  
(/u/5f50d2b82e05)

+ 关注

小礼物走一走，来简书关注我

赞赏支持



喜欢 (/sign\_in?utm\_source=desktop&utm\_medium=not-signed-in-like-button)

5



(http://cwb.assets.jianshu.io



3.jpg)

(/apps/download?utm\_source=nbc)



登录 (/sign\_in?utm\_source=desktop&utm\_medium=not-signed-in-comm

10条评论

只看作者

按喜欢排序 按时间正序 按时间倒序



九问的烦恼 (/u/320671844edb)  
2楼 · 2017.02.08 16:00  
(/u/320671844edb)  
博主，你用的时候没遇到bazel的问题么。。为什么我用的bazel各种问题。。好醉

赞 回复

ToeKnee (/u/5f50d2b82e05)： 啥问题？我这样装的没啥问题诶  
2017.02.23 22:22 回复

9e7b8eee8ccd (/u/9e7b8eee8ccd)： --input\_file\_pattern="{MSCOCO\_DIR}/train-?????-of-00256" \  
请问这一步是什么意思，这些问号就直接打上去？  
2017.04.22 09:31 回复

添加新评论



9e7b8eee8ccd (/u/9e7b8eee8ccd)  
3楼 · 2017.04.22 09:31  
(/u/9e7b8eee8ccd)  
--input\_file\_pattern="{MSCOCO\_DIR}/train-?????-of-00256" \  
请问这一步是什么意思，这些问号就直接打上去？


赞 回复


ToeKnee (/u/5f50d2b82e05)： 是的，那是个pattern  
2017.04.22 10:14 回复

Antigen (/u/95224ec365a8)： 博主能加个微信吗，最近在做这个作业啊





2017.05.11 21:24  回复

 添加新评论



lolipop\_3e49 (/u/74f978527175)

4楼 · 2017.07.28 13:19

(/u/74f978527175)

求教：可以利用作者已经训练好的CNN然后加到自己改良的LSTM吗

 赞  回复



offbye西涛 (/u/e0b1c6b05db0)

5楼 · 2017.09.12 15:13

(/u/e0b1c6b05db0)

这个模型 大概占用多少 磁盘空间？

 赞  回复



窗外篱笆小的小猫 (/u/f23971cd7127)

6楼 · 2017.10.26 20:35

(/u/f23971cd7127)

@9e7b8eee8ccd (/users/9e7b8eee8ccd) 请问你用这个--input\_file\_pattern="\${MSCOCO\_DIR}/train-?????-of-00256" \

有没有出现 Found no input files matching 上面的input\_file\_pattern呢？

 赞  回复



窗外篱笆小的小猫 (/u/f23971cd7127)

7楼 · 2017.10.26 20:36


(/u/f23971cd7127)

博主，请问你用这个--input\_file\_pattern="\${MSCOCO\_DIR}/train-?????-of-00256" \

有没有出现 Found no input files matching 上面的input\_file\_pattern呢？

 赞  回复



 十一维的风 (/c/6ce195da33a0?utm\_source=desktop&utm\_medium=notes-included-collection)

Machine... (/c/ec72d644eb23?utm\_source=desktop&utm\_medium=notes-included-collection)

[illegible]

几个月前，我写了一篇关于如何使用CNN（卷积神经网络）尤其是VGG16来分类图像的教程，该模型能够以很高的精确度识别我们日常生活中的1000种不同种类的物品。那时，模型还是和Keras包分开的，我们得从free-standing GitHub repo上下载并手动安装；现...

深度学习的需要了解的一些术语 ([/p/075f633a6913?utm\\_campaign=males...](http://075f633a6913?utm_campaign=males...))

御风之星 (/u/e6ae6d978f3d?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

图像识别是当今深度学习的主流应用，而Keras是入门最容易、使用最便捷的深度学习框架，所以搞图像识别，你也得强调速度，不能磨叽。本文让你在最短时间突破五个流行网络结构，迅速达到图像识别技术前沿。 作者|Adrian Rosebrock 译者|郭红广 编辑|鸽子 几个月...

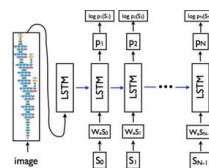



2017/11/30 下午5:57

utm\_medium=seo\_notes&utm\_source=recommendation)

解析：Google开源的“Show and Tell”，是如何让机器...

电影《HER》中的“萨曼莎”是一款基于AI的OS系统，基于对西奥多的手机信息和图像内容的理解，“她”可以为他处理日常事物、可以陪他谈心、甚至进行Virtual Sex，还可以读懂所有的书、跟哲学家交流，“她”所做的一切俨然就是一个有血有肉的人类才能实现的。但萨曼莎还胜于人...

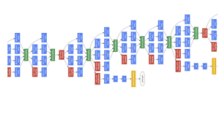



 图普科技 (/u/2c24cee41e7f?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/3f3b7e4d1281?utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

[MXnet] Simple Introduction to the Example (/p/3f3...

MXnet的学习笔记，这次主要是MXnet提供的example的综述介绍。关于MXnet在OSX下的编译安装，可以看这里Mac下编译安装MXNet!!!简介MXnet的样例程序分为5个部分，分别是DeepLearningExamples包括各种深度学习应用与比赛的实例Lan...




 ToeKnee (/u/5f50d2b82e05?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/5cf320bd138e?utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

据说近视眼的世界美得不像话？可现实是 (/p/5cf320bd...


据说近视眼的世界美得不像话？网友：近视眼的痛苦：三十米开外雌雄同体，五十米开外人畜不分，一百米开外六亲不认 网友：真得个近视眼折磨死你啊 站着说话不腰疼 网友：美个鬼！谁懂我看人看不见的感觉！世界都是模糊的！网友：你想多了，近视不戴眼镜只是满满的没有安全感 网友：近视眼...



 会奖旅游 (/u/76f29f387b42?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

2017-10-26 (/p/f25dda33c918?utm\_campaign=maleskine&utm\_content=...

没有一个女人，喜欢你讲道理，你讲的是道理，她听的却是你对她的态度。

 能不能陪着我 (/u/ff6acdc3581?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/b3821d4d8027?utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

[PHP文件管理器]①③--返回上一级 (/p/b3821d4d8027...


修改代码 index.php

返回上一级

ame(\$path)函数完成返回上一级

目录之后就不能在返回了




 子木同 (/u/53b741590a7e?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/75dbb32f99f1?utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

牛肉面馆这么经营，夏天你的生意比小龙虾还要好！（I...


餐饮吸引顾客进店就是这么简单 1、首先需要确认好自己的消费人群，一个最正常的普通人，消费水平一般，只是想吃一顿午饭。试想要是路过一家店面，如果装修的很豪华，看上去就很有档次，你就会考虑自己是否消费得起，大多数的人会选择绕过这家店，所以要把门面装修的简单一点，让一个消费者觉得...



 小百通 (/u/811fef2326b6?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

今天感悟 (/p/c040cd266f0f?utm\_campaign=maleskine&utm\_content=n...

人不要贪财，不要因为贪小便宜而忘了初衷。这样容易得不偿失。

 寒烟翠000 (/u/ea2ec17b9ec9?utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

